Andreas Kamilaris · Volker Wohlgemuth · Kostas Karatzas
Ioannis Athanasiadis

Editors

Preprint

Environmental Informatics

New perspectives in Environmental Information Systems: Transport, Sensors, Recycling

Adjunct Proceedings of the 34th edition of the EnviroInfo – the long standing and established international and interdisciplinary conference series on leading environmental information and communication technologies

Nicosia, Cyprus, September 23-24, 2020

EnviroInfo 2020 has been supported by

RISE UP

RESEARCH CENTRE ON INTERACTIVE MEDIA
SMART SYSTEMS AND EMERGING TECHNOLOGIES

Δήμος Λευκωσίας
Nicosia Municipality

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

EnviroInfo 2020 Organizers

General Chairs

Assist. Prof. Dr. Andreas Kamilaris, Team Leader at RISE, Cyprus

Prof. Dr. Volker Wohlgemuth, HTW Berlin, University of Applied Sciences, Berlin, Germany

Prof. Dr. Kostas Karatzas, Aristotle University of Thessaloniki, Greece

Assoc. Professor Dr. Ioannis Athanasiadis, Wageningen University & Research, Netherlands

Programme Committee

Antoniades, Demetris, RISE, Cyprus

Argyropoulos, Dimitrios, University College Dublin, Ireland

Arndt, Hans-Knud, Institut für Technische und Betriebliche Informationssysteme (ITI), Germany

Athanasiadis, Ioannis, Wageningen University & Research, Netherlands

Awad, Mariette, American University of Beirut, Lebanon

Bartoszczuk, Pawel, SGH Warsaw School of Economics, Poland

Behrens, Grit, University of Applied Sciences, Bielefeld, Germany

Canut, Carlos Granell, Universitat Jaume I., Spain

Castell, Núria, Norwegian Institute of Air Research (NILU), Norway

Charalambides, Alexandros, Klimate-KIC, Cyprus

Chatzichristofis, Savvas, Neapolis University Paphos, Cyprus

Cole, Ian, University of Cyprus, Cyprus

Constanti, Panayiota, Centre for Social Innovation, Cyprus

Engelhardt, Juri, ITC, University of Twene, Netherlands

Fakas, Georgios, Uppsala University, Sweden

Fishbain, Barak , Technion, Israel

Fountas, Spyros, Agricultural University of Athens, Greece

Fuchs-Kittowski, Frank, HTW Berlin, Germany

Geiger, Werner, Karlsruhe Institute of Technology, Germany

Greve, Klaus, University of Bonn, Germany

Guest, Olivia, RISE, Cyprus

Hadjisofocli, Demetris, Centre for Social Innovation, Cyprus

Hilty, Lorenz M., University of Zurich, Switzerland
Iliadis, Lazaros, Democritus University of Thrace, Greece
Intizar, Ali, Insight Centre for Data Analytics, Ireland
Jensen, Stefan, European Environment Agency (EEA), Denmark
Kalluri, Balaji, Technical University of Denmark, Denmark
Kamilaris, Andreas, RISE, Cyprus
Karantzalos, Konstantinos, National University of Athens, Greece
Karatsiolis, Savvas, RISE, Cyprus
Karatzas, Kostas, Aristotle University of Thessaloniki, Greece
Katos, Vassilis, Bournemouth University, UK
Khalifeh, Ala, German Jordanian University, Jordan
Knetsch, Gerlinde, German Environment Agency, Germany
Kolehmainen, Mikko, University of Eastern Finland and Forcenetics Oy, Finland
Kolios, Panayiotis, KIOS, Cyprus
Kompatsiaris, Ioannis, CERTH-ITI, Greece
Kondepudi, Sekhar Narayana, National University of Singapore, Singapore
Kotsev, Alexander, European Commission, Joint Research Centre (JRC), Belgium
Kranzlmüller, Dieter, Leibniz Supercomputing Centre, Germany
Kremers, Horst, CODATA, Germany
Lambrinos, Lambros, Cyprus University of Technology, Cyprus
Lanitis, Andreas, Cyprus University of Technology, Cyprus
Lestas, Marios, Frederick University, Cyprus
Liu, Lanfa, Institut Géographique National France, France
Loizos, Michael, Open University Cyprus, Cyprus
Loizou, Savvas G., Cyprus University of Technology, Cyprus
MacDonell, Margaret, Argonne National Laboratory, USA
Mashaly, Maggie, German University in Cairo, Egypt
Naumann, Stefan, Hochschule Trier, Umwelt-Campus Birkenfeld, Germany
Nikoletseas, Sotiris, Patras University, Greece
Oliver, Sergi Trilles, Universitat Jaume I., Spain
Osaragi, Toshihiro, Tokyo Institute of Technology, Japan
Ostermann, Frank, ITC, University of Twene, Netherlands
Otjacques, Benoît, Luxembourg Institute of Science and Technology, Luxembourg
Pitsillides, Andreas, University of Cyprus, Cyprus
Prenafeta, Francesc, Institute of Agrifood Research and Technology, Spain

Psara, Emily, Centre for Social Innovation, Cyprus

Savé, Robert, Institute of Agrifood Research and Technology, Spain

Sirmacek, Beril, Jönköping University, Sweden

Smith, Brendan, Insight Centre for Data Analytics, Ireland

Stütz, Peter, Bundeswehr University Munich, Germany

Themistokleous, Sotiris, Centre for Social Innovation, Cyprus

Thimm, Heiko Henning, Hochschule Pforzheim, Germany

Tsaltas, Dimitris, Cyprus University of Technology, EIT Food, Cyprus

Vassiliades, Vassilis, RISE, Cyprus

Vassiliou, Vasos, UCY/RISE, Cyprus

Voigt, Kristina, Helmholtz Zentrum München, Germany

Wagner vom Berg, Benjamin, University of Applied Science Bremerhaven, Germany

Weinberg, Volker, Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, Germany

Weismüller, Jens, Leibniz Supercomputing Centre, Germany

Willenbacher, Martina, HTW Berlin, Germany

Winter, Andreas, Carl von Ossietzky University Oldenburg, Germany

Wittmann, Jochen, HTW Berlin, Germany

Wohlgemuth, Volker, HTW Berlin, Germany

Zinonos, Zinon, Neapolis University Paphos, Cyprus

About the Editors

Andreas Kamilaris is a team leader at the Pervasive Real-World Computing for Sustainability (SuPerWorld) Multidisciplinary Research Group (MRG) of the newly established Research Centre on Interactive Media, Smart Systems and Emerging Technologies (RISE). In parallel, he is an Assistant Professor at the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) of the University of Twente. His research interests are Internet/Web of Things, geospatial analysis, pervasive computing, smart environments and machine learning.

Volker Wohlgemuth is a Professor for Industrial Environmental Informatics at the Faculty of Engineering – Technology and Life, HTW Berlin, University of Applied Sciences. His research fields are Material Flow Management, Modeling and Simulation, Industrial Symbiosis and Environmental Management Information Systems.

Kostas Karatzas is a Professor at the School of Mechanical Engineering, Aristotle University of Thessaloniki, leading the Environmental Informatics Research Group. Kostas does research in Environmental Informatics and Modelling, Mechanical Engineering and Computational Intelligence.

Ioannis Athanasiadis is an Associate Professor in Data Science with the Laboratory of Geoinformation Science and Remote Sensing at Wageningen University, Netherlands. His expertise includes data science, big data, environmental informatics, software engineering and intelligent information systems.

Preface

This book presents short papers and work in progress papers of the 34th edition of the long-standing and established international and interdisciplinary conference series on environmental information and communication technologies (EnvirolInfo 2020).

The conference was held from 23 –24 September 2020 virtually. It was organized by the Research Centre on Interactive Media, Smart Systems and Emerging Technologies (RISE), Nicosia, Cyprus under the patronage of the Technical Committee on Environmental Informatics of the Gesellschaft für Informatik e.V. (German Informatics Society – GI). RISE is a research centre of excellence in Cyprus, aiming to empower knowledge and technology transfer in the region of South-East Mediterranean. It is a joint venture between the three public universities of Cyprus (University of Cyprus, Cyprus University of Technology and Open University of Cyprus), the Municipality of Nicosia, and two renowned international partners, the Max Planck Institute for Informatics, Germany, and, the University College London, United Kingdom.

Combining and shaping national and international activities in the field of applied informatics and environmental informatics, the EnviroInfo conference series aims at presenting and discussing the latest state-of-the-art development on information and communication technology (ICT) and environmental related fields. A special focus of the conference was on digital twins and, in particular, the emerging research concept of digital twins for sustainability, where natural systems are twinned with digital replicas, to improve our understanding of complex socio-environmental systems through advanced intelligence. Sustainable digital twins of smart environments are also a flagship project of RISE.

This paper collection covers a broad range of scientific aspects including advances in core environmental informatics-related technologies, such as earth observation, environmental monitoring and modelling, big data and machine learning, robotics, smart agriculture and food solutions, renewable energy-based solutions, optimization of infrastructures, sustainable industrial/production processes, and citizen science, as well as applications of ICT solutions intended to support societal transformation processes toward the more sustainable management of resource use, transportation and energy supplies.

We would like to thank all contributors for their submissions. Special thanks also go to the members of the programme and organizing committees, for reviewing all submissions. In particular, we like to thank our local organizers at RISE who responded fast and generated a digital twin of the physical conference and hosted it online. We also deeply appreciate the help and support of the Environmental Informatics community that backed up our efforts to cope with

the COVID-19 pandemic and to have a stimulating and productive online event. Last, but not least, a warm thank you to our sponsors that supported the conference.

Andreas Kamilaris Nicosia, Cyprus
Volker Wohlgemuth, Berlin, Germany
Kostas Karatzas, Thessaloniki, Greece
Ioannis N. Athanasiadis, Wageningen, The Netherlands

Nicosia, December 2020

Table of Contents

PART I
TRANSPORT, MOBILITY AND LOGISTICS

Improving Delay Forecasts in Public Transport using Machine Learning Techniques

Henning Wobken[1], Alexander Dölling, Jon-Patric Ewelt, Niklas Howad, Florian Hustede, Hendrik Jordan, Abdalaziz Obead, Jari Radler, Sebastian Schnieder, Klaas von der Heide, Ole Wehrmeyer, Mathias Wille, Barbara Rapp, Jorge Marx Gómez

1. Introduction

Due to the increasingly intense debate on climate change in recent years, local public transport is once again assuming an increasingly central role in the discussion on mobility concepts to substitute the focus private transportation used to have in the last decades [16]. Improving the attractivity of public bus transport is a key factor in making it competitive against car-based mobility concepts, e.g. carpooling, car sharing, and depends on the reliability and punctuality [7] of the service provided. To achieve this, the public transport service must deliver accurate information about the service of lines to its customers, enabling them to adjust their schedules or look for alternatives.

To improve the attractiveness of public transport, it is crucial to improve the information a public transport provider can relay to their customers, especially the quality of the provided delay prediction [9]. To improve the prediction, first, the factors which influence the delay of public transport services, e.g. weather and traffic, must be evaluated. Second, it must be determined how severe the influence of these factors is. To evaluate the factors influencing public transport service and to train a machine learning algorithm, this paper uses data provided by transportation providers. These providers are obligated by law to track their busses and store data about their service performance. In cooperation with AMCON, a German developer of bus data information systems [5], we are able to train a machine learning algorithm with historic service data from two transportation providers in rural and mid-level city environments. Weather and traffic data are retrieved from available interfaces.

2. Related Work

Providing accurate information on arrival and departure times in public transport is one of the key parameters for high-quality public transport. Many studies have been conducted to assess the accuracy of predictions using different data sources, methods and models. The data for the development of such models can have different sources and can either be collected historically or be available in real time: Automatic Passenger Counting (APC), Automatic Vehicle Location (AVL) and Global Positioning Systems (GPS). There are different approaches in the literature to classify the mathematical methods and models. It is possible to classify the proposed models into four categories [19]: Regression models, models for artificial neural networks (ANN), Kalman filter models and analytical approaches. In another comparison, the

[1] Carl von Ossietzky Universität Oldenburg, 26129 Oldenburg, Germany, henning.wobken@uni-oldenburg.de

classification of analytical approaches is omitted, but the categories support vector models and nearest neighbor methods are added [4]. Another approach using a log-normal auto-regressive (AR) model approach has been successfully implemented [14] and shows that the development of a suitable model is far from complete and the classification of the models is still in discussion.

Using a regression model, correlations between several characteristics are represented by a mathematical model. A basic distinction is made between dependent and independent variables. To determine the bus travel time, different variables such as traffic flow at intersections, weather conditions or passenger demand at bus stops are decisive influencing variables [11]. Using regression models, bus travel times have been accurately predicted both under normal traffic conditions and during a temporary road closure, e.g. due to a road improvement measure [2]. In a research on the prediction of bus arrival times at bus stops with several routes, a linear regression model was successfully applied among other methods, but it turned out that in this case a Support Vector Machine (SVM) model, a statistical approach to the classification of objects, performs best for the prediction among four proposed models [30].

Another tool is the Kalman filter, which is a mathematical model for the iterative estimation of parameters based on faulty observations within a system. With the help of this filter technique, Wall and Dailey were already able to set up a first algorithm in 1999 which, in combination with GPS data and historical data, tracked the locations of public transport vehicles to predict travel times [29].

Artificial neural networks (ANN) are another approach to delay prediction, training a system inspired by the connectivity seen in the brain. Despite the theoretical background of ANN dating back to the early twentieth century [20], applying ANN especially in solving delay prediction in public transport is a novel concept [22]. To predict the delay for railway services in Germany [23], a neural network was trained and then evaluated against a rule-based system which factors in experiences and historic data. The rule-based system will use predetermined delay scenarios with the expectation that the neural network is able to abstract from known constellations causing delay, which was concluded to be the case.

Besides the actual model classification, there are many different approaches to develop dynamic arrival time prediction models, as described in a Google AI blog post [15]. In this article, the developers at Google describe the efforts the company took to improve the prediction data of the bus delays to use in their geodata service "Google Maps". The approach is splitting the bus route into multiple parts, each one gets its own delay prediction based on traffic data, and summing up each part for a total prediction on the whole route. The underlying algorithm is called long short-term memory (LSTM) and describes a special function block of recurrent neural networks, through which a kind of "long short-term memory" can be integrated [17]. The data source is provided by information the company collects from their users, used to train their machine learning algorithm. Furthermore, the discussion they conducted with members of their “Google Maps”-Team coincides with our own considerations on the issue of influence factors for bus transportation delay in general.

3. Status Quo

To illustrate the status quo, a brief introduction to the organizational structures of the German public transport system is required. In 2018, there were 2.208 bus companies operating scheduled public transport services in Germany [9]. The line networks are managed by transport associations, of which there were

648 for public transport in 2018 [27, 24]. However, these associations do not cover the entire area of Germany. In those areas without an association, the networks are managed by the district in cooperation with bus companies. Due to the complex structure of German public transport, several different bus companies can operate in a city or region. The passenger has the requirement that the timetables can be viewed centrally. Furthermore, there are apps and displays at the stops that show the planned arrival times of the buses. The buses can now be operated by different companies and use different independent software systems. This requires central data hubs to which all systems report the current timetable situation so that end customer apps can query them.

These central data hubs require standardizations for the transmission of data. These are adopted and administered by the "Verbund Deutscher Verkehrsunternehmen" (VDV). This is an industry association of the public transport sector, comprising transport companies, transport associations, clients and manufacturers [27, 24]. The VDV consists of 450 transport companies that handle 90% of the total volume of public transport in Germany. The adopted standards are called VDV writings. A list of all standards is published on the website of the VDV [28].

The central data hubs use VDV453 [25] and VDV454 [26]. These standards indicate that the target and actual timetable data are sent by the respective software system of the bus enterprise. The forecasts for arrivals at the stops are thus calculated by different systems and can also be of varying quality. The central data hub only receives this data and makes it available to other services, e.g. an end customer app. Against this background, the optimization of the forecast takes place in the context of one bus operator.

Based on an expert interview with Olaf Clausen, Managing Director of AMCON, we collected information about how predictions are determined in the current system and which data sources are used for this prediction. The data basis is the target timetable, which contains all trips and stops with GPS coordinates. Added to this, the buses report their position every ten seconds and whether they have left a stop. The current forecast of arrivals and departures is based on a simple algorithm. When a bus starts its journey, the predicted arrival times are set equal to the target timetable. At each departure at a stop, the arrival times for the following stops are recalculated. The basis for this calculation is the deviation from the target timetable. If, as an example, a bus departs with one-minute delay, a delay of one minute is set for the following stops.

4. Data Source

This section explains our data procurement and collection. This data is prepared in a self-modelled data science process using the data warehouse reference architecture of Bauer and Günzel [6]. For this purpose, the data was divided into four dimensions (actual arrival data, target arrival data, traffic and weather) of a star schema and merged into a fact table. Based on this fact table the models could be trained. Figure 1 shows the data science process. Various data sources for weather, traffic and the bus operator's data are examined. ETL processes integrate the data into the data warehouse. The data is analyzed to determine various factors for the forecast. Training and test data are extracted and applied to machine learning models. However, only part of the possible data for machine learning is extracted to test the models for unknown data. Finally, the predictions of the models are evaluated using different metrics, results of other algorithms and the actual delays.

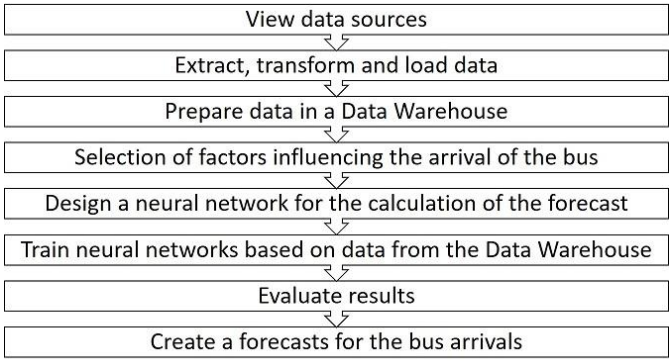


Fig. 1. Processes of data preparation and model generation.

The fact table has a data set of more than five million rows with various columns for dimensions, for example jam-factor, temperature, precipitation and datetime. Weather data is provided for the entire data set, but only 6,536 rows contain information on traffic flow. Therefore, the influence of traffic on bus arrival cannot be considered in its entirety. Different columns from the fact table were discussed for using them as features for the MLP models. The analysis of [19] was also considered. Time, day of the week, month, precipitation, traffic and temperature were defined as influencing factors on a bus arrival time.

These features were validated by correlation analysis and the use of an MLP model with one feature each. The results of the analysis are shown in Table 1. We evaluate the traffic, month and day of the week as a negative factor on the arrival of the buses, because a traffic jam on a road leads to delays and there are days and months when a bus is used more often as a means of transport. Temperature and time have a positive influence as there are certain times when buses are used less, and rising temperature can lead to more use of other modes of transport. We see the precipitation as a negative influence, but we cannot prove this with our data.

column	time	weekday	month	precipitation	traffic	temperature
correlation	0.137897	-0.059135	-0.529300	0.008811	-0.016876	0.290177

Table 1. Correlation analysis of the bus delay.

The correlations can only be used as decision support for the selection of factors, since the correlation analysis did not yield clear results for one or more factors. This problem is caused by partially missing weather and traffic data. Nevertheless, positive or negative results can be interpreted for the influence of the factors on bus travel. Based on this data and selected attributes, different artificial intelligence methods were evaluated.

5. Artificial Intelligence

As shown in section 2, there have been numerous approaches, each with varying amount of success. In this chapter, we discuss different approaches and determine a suitable one to be used in the model implementation. AI is a broad field, which generally deals with enabling a computer to solve tasks which require intelligence [12]. Since we want to use the available data to automatically train a model, we need to enter the field of machine learning, which deals with self-learning artificial intelligence systems [18]. A bus arrival delay can be any real number within a reasonable time frame. This type of problem requires a numerical, continuous function to be solved. A regression under a supervised learning procedure can be used [8, 18].

There are several algorithms which can be used for regression. Because some algorithms can provide more accurate predictions than others regarding a specific problem, a suitable algorithm is determined by recommendations found in the literature and tests conducted on an early version of our data warehouse. For the tests, a model was implemented for each algorithm using Tensorflow with Keras or scikit-learn. The following four algorithms were considered:

Multilayer Perceptron (MLP) is a type of ANN. It can learn the pattern and presentation of training data, enabling it to make suitable predictions even for complex problems [3].

Support Vector Regression (SVR) is the adaption of an SVM to a regression problem [21]. Since input data of the problems discussed in this paper are not linearly separable, the SVR which would have to be used could be a two-layer neural network [18].

Decision Tree While Decision Trees are commonly used for classifications, they can also be used for regression [13]. Using the Random Forest algorithm, a more complex tree consisting of multiple decision trees can be formed for more accurate predictions [1].

Long short-term memory (LSTM) is a type of RNN that was introduced by Hochreiter and Schmidhuber in 1997 [17]. It wants to avoid the error back-flow problems of other RNNs by using gradient based learning algorithms [17]**.**

Delays in public transport can be dependent on many different factors, which might form complex relationships. An MLP model is well suited for recognizing these complex relationships [18]. In accordance with this, our tests showed that measured by the metrics R^2 and mean absolute error (MAE), the MLP model would give the best results. Therefore, MLP will be used in the model implementation. However, it must be noted that the focus of these tests was to determine a suitable algorithm quickly and not to achieve the best possible result for each algorithm. Therefore, we considered the long short-term memory approach as an additional way to make more accurate predictions for a stop. For the testing of MLP we used Tensorflow with Keras, which will also be used in the model implementation.

6. Model implementation

This section describes the process of implementing machine learning models for forecasting bus arrivals. In our work we used two different approaches one MLP model and one LSTM network.

In the determination and implementation process of the MLP model the first step is the determination of input factors, the number of layers and neurons and activation and optimizer functions. Our

implementation of the genetic algorithm trains 15 models in 20 generations and uses the R^2 as the fitness metric. The first generation are 20 models whose parameters have been chosen randomly from a list of given choices. Table 2 shows the available values for our parameters.

Hidden Layers:	**1, 2, 3, 4, 5**
Neurons:	4, 8, 16, 32, 64, 128, 256, 512, 768, 1024
Optimizers:	adam, RMSprop, sgd, adagrad, adadelta, adamax, nadam
Activations:	relu, tanh, sigmoid, elu

Table 2. Parameters for the evaluation of the model.

After each generation the models are sorted based on their R^2 score with 1 being the best possible score. The top 40% of each generation are kept for the following generation without changing them. Of those last 60% with lower scores, 10% are also randomly kept. Those two groups make up 50% of the next generation. The other 50% are children bred from the top 40% of the previous generation. The best model of the final generation has the following parameters: hidden layers 4, neurons 768, optimizer adamax and activation elu.

The generated model was trained on a database of 2,400,000 entries of 2018 and 2019 and tested with 600,000 entries. This model has a value of $R^2 = 0.69$. The average forecasted delay was 3690.1 seconds whereas the real delay averages at around 55.76 seconds.

The average values and deviations show that the models do not recognize the actual delays. Our next step was the introduction of new input parameters. We concluded that the lacked information about how long the distance between the bus stop is, and how much of a delay there was on the previous bus stop of the line. For these new input values, the correlation values were calculated in order to obtain a comparison in the correlation between delay and input value. Those new input values have a correlation of 0.82 for the previous delay and -0.22 for the distance to the next stop.

After the calculation and introduction of those new parameters we started a second run of the genetic algorithm with 50 models in 30 generations. The model has the following parameters: hidden layers 5, neurons 128, optimizer nadam and activation elu. It was also trained on 600,000 entries of our dataset.

The resulting model has a value of $R^2 = 0.83$. The average prediction changed to -6.48 seconds and therefore has only a deviation of 62.24 seconds from the actual delays. This allowed us to determine that the previous delay is a crucial factor in achieving more efficient model. In addition, the distance has an influence on the determination of the delay, since we could only improve the model with both new input values.

As the current results are not yet satisfactory, a different approach is being evaluated. The idea is not to consider the arrivals at bus stops, but to split a bus route into a list of segments and calculate the time needed to travel in each segment. The estimate time of arrival at a bus station can then be obtained by accumulating the travel times for all segments between the current bus position and the target bus stop.

To achieve this, at first, all bus routes were split into shared segments, with the goal that each segment is both as long as possible (but never longer than the route between two stops) and shared between as many bus routes as possible. This process yielded 3,518 route segments for the analyzed area, resulting in as many models that need to be trained. An example for segments between bus stops is shown in Figure 2.

Fig. 2. Example of route segmentation between two bus stops.

For the preparation of the training, the time to travel the segment at a certain point in time needs to be calculated. This is being done by gathering the raw GPS data and timestamps of buses travelling there and calculating the travel time. With 170 million raw GPS datapoints, this results in around 41,000 points per segment, assuming an even distribution.

With these data, different types of models will be trained and tested. We have high hopes, that a LSTM model will perform very well in this scenario, as the memory function can for example recognize a deteriorating road state or deal better with changing weather conditions. We also hope that the increasing specialization for the model(s) can more accurately predict the conditions, and thus needed time, that the bus will have to face on that part of the road. The parameters that are going to be used for training of the models for each segment are going to be Time of day, Weekday, Month, Precipitation and Temperature.

As there are still few reliable traffic data, these are not going to be included in the model at all. In theory, the major rush-hour traffic on main roads should be accurately reflected in the historical data. The lack of traffic data also should be less of a problem for this approach due to the much higher geospatial resolution of the space the model is used for. Training one general model to factor in rush hour between certain stops of certain routes is far more difficult than training it only for the specific models of the road segments affected by rush hour.

7. Conclusion and Outlook

In this paper, bus positions and timetables data of bus provider companies in Germany were combined in a data warehouse with data from external services regarding weather and traffic. Based on these data, different machine learning models were trained using the MLP model approach. Metrics like R^2 and mean average error of the models were compared. The comparisons have shown that an MLP model can make a prediction of the delay and that there is only a small deviation from reality.

While our MLP approach involved training one generalized model, our LSTM approach involves training many specialized models. While the route segmentation needed for training the models is already finished, our LSTM models are still under development. Therefore, we could not yet present any

meaningful results. But we expect the LSTM models to adapt better to the conditions of specific segments in the route network. Additionally, in the future a metamodel integrating both the LSTM model with the route segment approach and the MLP model can be tested as well. This might combine the best of the two quite different approaches.

References

[1] Abbott, D.: Applied Predictive Analytics. Wiley (2014)

[2] Abdelfattah, A.M., Khan, A.M.: Models for Predicting Bus Delays. Transportation Research Record, (vol. 1623), 8-15 (1998). https://doi.org/10.3141/1623-02

[3] Alpaydin, E.: Introduction to Machine Learning. The MIT Press Cambridge, Massachusetts, third edition edn. (2014)

[4] Altinkaya, M., Zontul, M.: Urban Bus Arrival Time Prediction: A Review of Computational Models. International Journal of Recent Technology and Engineering (vol. 2), 164-169 (2013)

[5] AMCON Software GmbH: Uber uns. https://amcongmbh.de/ueber-amcon.html (2020)

[6] Bauer, A., Gunzel, H. (eds.): Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung. dpunkt.verlag, Heidelberg, 4., uberarbeitete und erweiterte auflage edn. (2013)

[7] Boltze, M., Specht, G., Friedrich, D., Figur, A.: Grundlagen fur die Beeinflussung des individuellen Verkehrsmittelwahlverhaltens durch Direktmarketing. Tech. rep., Darmstadt Technical University, Department of Business Administration, Economics and Law, Institute for Business Studies (BWL) (2002)

[8] Brownlee, J.: Master Machine Learning Algorithms. v1.1 edn. (2016)

[9] Bundesverband Deutscher Omnibusunternehmer e.V.: Zahlen, Fakten, Positionen. https://www.bdo.org/zahlen-fakten-positionen (2020)

[10] Ceder, A.: Public Transit Planning and Operation, Theory, Modelling and Practice. Elsevier, Oxford (2007)

[11] Chen, M., Liu, X., Xia, J., Chien, S.I.: A Dynamic Bus-Arrival Time Prediction Model Based on APC Data. Computer Aided Civil and Infrastructure Engineering (vol. 19), 364-376 (2004)

[12] Coppin, B.: Articial Intelligence Illuminated. Jones and Bartlett Publishers, Boston, 1st ed edn. (2004)

[13] Denison, D.G.T. (ed.): Bayesian Methods for Nonlinear Classication and Regression. Wiley Series in Probability and Statistics, Wiley, Chichester, England; New York, NY (2002)

[14] Dhivya Bharathi, B., Anil Kumar, B., Achar, A., Vanajakshi, L.: Bustravel time prediction: A log-normal auto-regressive (AR) modelling approach. Transportmetrica A: TransportScience (vol. 10), 807-839 (2020). https://doi.org/10.1080/23249935.2020.1720864

[15] Fabrikant, A.: Predicting Bus Delays with Machine Learning (2019)

[16] Hayashi, Y., Matsuoka, I., Fujisaki, K., Itoh, R., Kato, H., Rothengatter, W., Takeshita, H.: Importance of intercity passenger transport for climate change issues. In: Hayashi, Y., Morichi, S., Oum, T.H., Rothengatter, W. (eds.) Intercity Transport and Climate Change: Strategies for Reducing the Carbon Footprint, 1-30. Springer International Publishing, Cham (2015)

[17] Hochreiter, S. & Schmidhuber, J.: Long Short-term Memory. Neural computation. 9. 1735-80. (1997) https://doi.org/10.1162/neco.1997.9.8.1735

[18] Khan, S., Rahmani, H., Shah, S.A.A., Bennamoun, M.: A Guide to Convolutional Neural Networks for Computer Vision (2018)

[19] Mazloumi, E., Currie, G., Rose, G., Sarvi, M.: Using SCATS data to predict bus travel time. In: Australasian Transport Research Forum 2009. 1-14. Australasian Transport Research Forum, Auckland New Zealand (2009)

[20] McCulloch, W.S., Pitts, W.H.: A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics (vol. 5), 115-133 (1943). https://doi.org/10.1007/BF02478259

[21] Mechelli, A. (ed.): Machine Learning: Methods and Applications to Brain Disorders. Elsevier, San Deigo, first edn. (2019)

[22] Pekel, E., Kara, S.S.: A Comprehensive review for artificial neural network application to public transportation. Sigma Journal of Engineering and Natural Sciences (vol. 35), 157-179 (2017)

[23] Peters, J., Emig, B., Jung, M., Schmidt, S.: Prediction of Delays in Public Transportation using Neural Networks. In: International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents. 92-97. Web Technologies and Internet Commerce (CIMCA-IAWTIC'06, Vienna (2005)

[24] Reinhardt, W.: Öffentlicher Personennahverkehr: Technik - Rechts- Und Betriebswirtschaftliche Grundlagen. Springer, Wiesbaden, second edn. (2018)

[25] Verband Deutscher Verkehrsunternehmen (VDV): VDV-Schrift 453 Ist-Daten-Schnittstelle (2018)

[26] Verband Deutscher Verkehrsunternehmen (VDV): VDV-Schrift 454 Ist-Daten-Schnittstelle - Fahrplanauskunft (2018)

[27] Verband Deutscher Verkehrsunternehmen (VDV): Daten & Fakten zum Personenund Schienengasterverkehr. https://www.vdv.de/daten-fakten.aspx (2019)

[28] Verband Deutscher Verkehrsunternehmen (VDV): Publikationsverzeichnis. https://www.vdv.de/vdv-publikationsverzeichnis-d.pdfx (2019)

[29] Wall, Z., Dailey, D.J.: An Algorithm for Predicting the Arrival Time of Mass Transit Vehicles Using Automatic Vehicle Location Data. In: Annual Meeting of the Transportation Research Board, National Research Council. 78th Annual Meeting, National Research, Washington D.C (1999)

[30] Yu, B., Lam, W.H.K., Tam, M.L.: Bus arrival time prediction at bus stop with multiple routes. Transportation Research Part C: Emerging Technologies (vol. 19), 1157-1170 (2011)

Decentralized Identity Management for DLT-based Cooperation Support

Enabling New Business Relationships for Inland Waterway Transport

Thomas Osterland[1] & Thomas Rose[1&2]

1. Introduction

Distributed ledger technology (DLT) allows one to securely store information and execute program code within a network. It generates trust algorithmically between the network participants. For that, it relies on a consensus algorithm, that ensures that all network participants have the same information. There is no central entity in place that synchronizes the information. The network synchronizes itself by means of consensus building.

Due to the particular characteristics of the distributed ledger technology no single entity is able to alter data as well as manipulate or prevent the execution of program code, once persisted on the distributed ledger. This protects the data in the sense, that data stored on the ledger cannot be altered. However, information on the ledger cannot per se be considered trustworthy or correct in the context of the application domain. Any temperature stored on the distributed ledger does not necessarily correspond to the actual weather outside. Instead, the authenticity and source of the information ensures its correctness, as well.

Originally, this concept for distribution has been applied merely for transaction management in the realms of financial services. Afterwards people realized that many transactions across business partners in various sectors of industry can be governed by DLT, that is, DLT matured to become an intermediary, enabling new business models and use cases. Collaboration across businesses is no more limited to established networks with their relationships or intermediaries as governance, but partners can establish a business relationship on the basis of DLT. Hence, collaborations become more flexible and open for partnerships on the fly.

Unfortunately, DLT suffers on the same problem as the internet: users are anonymous and we do not know with whom we communicate and cooperate. In the age of the Internet of Things (IoT), it is even possible that our communication partner is a sensor or a machine that we do not know. Moreover, any partner for process collaboration can merely be technical agents with some believes and desires.

Besides the potential namelessness of business partners, assessment and assurance of information quality comes as critical challenge on top. Once the trustworthiness of a business cooperation is founded in information, integrity of this information is essential and specific knowledge about the identity of our communication partner delivering information is compulsory. By “specific” we refer to a particular view on the identity. Since we do not necessarily need to know the clear name and address or the social security number, but we need to have means to verify concrete attributes attached to the identity of a person. Imagine, you want to buy alcoholic beverages from the liquor store: it merely matters that you cover the

[1] Fraunhofer FIT, Schloss Birlinghoven, 53754 Sankt Augustin, Germany, thomas.osterland@fit.fraunhofer.de
[2] RWTH Aachen University, Informatik 5, Ahornstraße 55, 52056 Aachen, Germany, thomas.rose@fit.fraunhofer.de

requirements of the legal drinking age. Hence, instead of providing a number or a birth date we talk about returning a Boolean value and thus, minimize the exchange of personal information.

Going down this avenue, one question arises: does it make sense to have a central authority that manages the attributes attached to personal information of an entity verifiable as we are currently used to? Such a centralized management of identities appears absurd. On the one hand, we promote the use of distributed ledger technology to foster decentral applications, or respectively enable such a distribution in first place, while on the other hand we favor a central authority to ensure the identity of network participants.

Besides undermining the security and the decentral nature of the distributed ledger technology, such a centralization of identity management leads to new questions and problems, as for instance about the governance and the operation of the central identity management platform from an economic perspective.

Research and industry responded to this problem with the proposal of standards for decentralized identity management, as for instance Verifiable Credentials (VC) [3] and Decentralized Identifiers (DID) [4]. These standards describe procedures and architectural models to implement decentral identity management without central authorities. A survey of existing implementations is given in [2] with a more thorough description of the popular Ethereum based decentral identity management solution uPort in [5]. Although these approaches and the aforementioned standards particularly feature decentralized identity management, the fundamental requirements on general identity management, as summarized by [1], are thoroughly considered as functional requirements on DID implementations.

In this paper, we discuss the use of decentral identity management in the context of a research project, that fosters the digitization of document exchanges in the domain of inland waterway transportation. Different stakeholders have to exchange information in order to manage the transportation of goods. The value of the information is based on the authenticity of the information. Thus,stakeholders are required to authenticate, when exchanging information. Even though the research project is only nationwide, the underlying logistic can hardly be limited to national borders. Hence, we need to verify the authenticity of individuals and organizations in different countries. Instead of implementing a central identity management, that is accepted by all participating countries, decentral identity management will be a solution to mitigate negotiation efforts beforehand in order to agree among many countries in the world, when considering global logistics.

Moreover, we also need means to govern the accessibility of process information that occur during the transportation of goods on a waterway. In the following, we discuss an approach based on attribute-based encryption in combination with decentral identity management to fathom role-based points of view on transportation information. Hence, visibility of information is tailored with regard to verified identities.

2. The Use Case of Inland Waterway Transportation

Logistics is decisive for the success of many business operations. The question arises of how to improve sustainability of transportation processes. Inland waterway transportation is one element in intermodal logistic processes. Although the volume of transport has roughly doubled over the last two decades, the share of waterway transportation has dropped by almost 20%. This means, that inland waterway transportation is an economic looser of this competition, although offering transportation services with

minimal resource consumptions. The question arises, how to make inland waterway transportation more competitive. Digitalization and DLT are considered as savior and business enabler for more flexible transportation networks. Both call for the identification of participating parties and information.

German and also European waterway transportation builds upon a well-developed network of waterways for the transport of goods with a dense network of inland ports. Goods transported are typically piece goods, containers or bulk loads. The latter are goods, as coal for power plants, sand and gravel for the construction industry and grain or animal nutrition. Liquid bulk, as chemicals or propellants are transported on the waterways, as well. They present the majority of goods transported on waterways, while containers have major shares on transport highways such as the Rhine. Piece goods are for instance large wire coils or cargo containers, that are transported from seaports to inland ports and vice versa. The transportation is handled by inland navigation operators, which can be self-employed individuals, that often live on their ships, but also shipping companies, that own a fleet of barges or cooperatives of inland navigation operators.

As of today, processing of transport orders is mostly paper-based with occasional oral agreements. Before and during the transportation, participating stakeholders exchange documents, that are digitized partially by stakeholders.

Hence, one aim of the research project is the digitization of paper-based document interchanges. Instead, we fathom the exchange of digital information, that can be automatically parsed and processed by computer systems and as such, provide digital benefits, as the simplification and acceleration of existing business processes, to all participating stakeholders. In addition, the research project aims at the assessment of the cooperation benefits provided by the distributed ledger technology employed in the domain of inland waterway transportation. In particular, we are interested in the potential automation and optimization of manual labor required in intermediate steps, for example, by introducing gauged sensors and IoT networks. This lays the groundwork for a more distant vision of a machine-to-machine economy, in which autonomous systems automatically cooperate in the real world, such as an autonomous car that is sent to refill petrol and thus, needs to communicate with the petrol station and handle economical demands, after it drives its owner to the local shopping center.

Exemplarily, we consider a specific use case in the domain of the German inland waterway transportation, that describes the transport of bulk good and focuses in particular on the loading and unloading procedures. The process starts with the issuing of the transportation order to the inland navigation operator. The operator approaches the loading port, which will inspect the cargo space before loading the goods. Depending on the transported goods, as for instance food or animal nutrition, the cargo space must be cleaned accordingly and the goods that were transported beforehand are of importance. So, it must be ensured that for instance grain is not mixed up with fertilizer. The result of the inspection is a document called FRI/CLI report, which certifies the qualification of the cargo room to transport the respective good.

The cargo space inspection is followed by the gauge. That is, the water displacement of the empty barge is measured. Although, there exists sensors, that can do the measurement for official reasons a gauger needs to read the displacement from a scale that is attached to the side of the barge. Hence, a kind of physical yardstick can be replaced by an electronic device. Then the loading process starts and is concluding with once again taking the gauge. The difference in the water displacement represents the weight of the loaded freight. Thereupon, the inland navigation operator obtains the bill of lading.

The inland navigation operator drives to the port of destination, where, first the freight is inspected. Then the water displacement is measured before and after unloading in order to determine the delivered weight of the freight. Transport is concluded by providing the operator with a certificate of unloading.

From the description of this use case we can learn a number of interesting challenges with regard to identity management. Different stakeholders exist, which are either part of an organization or individuals. For instance, a shipper in the port is part of the port organization and might sign documents as authority of the port. A gauger, which is an official appointed by bylaws, however, needs to sign with an individual identity, that refers to a commission. However, the gauger is also a representative of a company. On the other way not, every employee of a company is entitled to determine the gauge.

3. Decentral Identity Management in Inland Waterway Transportation

The design of a decentral identity management for the German inland waterway transportation is strongly related to the roles of participating parties in the underlying logistic process. The waterway transportation use case, that we discussed in the previous section, showcases a variety of requirements to adept an identity management system that is accessible and comprehensible to all affected parties.

Considering a completely digitized and distributed ledger-enabled transportation process, we figure that the transportation order will be encoded into a smart contract. The operator and organizer of a transportation order instantiates a smart contract into a public distributed ledger, that holds all the important basic conditions, as for instance place of dispatch, destination, estimated time of arrival, freight class and prices. One could even fathom the implementation of exception handling and automatic penalties, if one party fails in providing demanded services with respect to the defined conditions. This elevates the smart contract into a full-fledged decentral autonomous organization (DAO); an ad-hoc business cooperation with the sole aim to provide the transportation service.

The public distributed ledger with the instantiated smart contract serves as a market place, where different inland navigation operators or shippers can bid on the contract. Ideally, a self-employed navigation operator has an expert system that automatically crawls the open contracts and proposes those that suit the current state of the ship and travel route.

This first step, requires comprehensive knowledge about the identities and roles of the involved parties. The identity of the creator of the transportation order must be verifiable, since the shipper must be sure that the order is serious, otherwise he loses the costs for stopping at the place of dispatch. Even though, one could encode a related penalty into the smart contract, a completely anonymous procurement is not usual. When we consider the inland navigation operator, the identity and attached roles become even more important: So, the smart contract should only consider inland navigation operators that have a valid driver license and insurance policy. In case the contract specifies the transportation of hazardous goods, the operator must dispose of a corresponding training. But also, the barge of the inland navigation operator has an identity and attached roles that are important to the transportation order encoded in the smart contract. So, the ship must have a valid operating license, adequate safety equipment, in case of hazardous goods it needs the corresponding facilities and gauged instruments.

The digital and decentral nature of an identity management that operates on top of a distributed ledger allows the smart contract to automatically validate the suitability of applicants and their equipment.

With certificates based on cryptographic methods it is even possible to model more complex identities and role attributes that are attached to it. So, for instance, it is possible to create certificate hierarchies, that allow one to identify an inland navigation officer to be part of a cooperative association, that might have special agreements. In particular, there might be relations between identities and aside from persons and organizations, identities can be assigned to machines or sensors. For instance, we can imagine that not only a ship has an identity, but also the different cargo holds of the ship. In this way different cargo holds can have different cleaning states that were digitally certified by corresponding CRI/LRI assessments. But also, sensory equipment might have identities. This allows the introduction of gauged displacement sensors that are tested regularly and replace the manual labor of taking the gauge.

Throughout the discussed transportation process there exists different stakeholders, namely shippers, gaugers, cargo hold inspectors, inland navigation operators and carrier. All these parties need to be served with reliable identities.

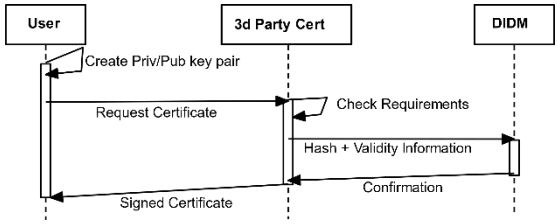


Fig. 1. Certification sequence of a user.

The sequence diagram in Figure 1 demonstrates the process of creating and qualifying identities. In a first step, a user is required to create a public/ private key pair. This is followed by applying to issuing third parties for certification. A certification process can have different forms and relies strongly on the context. So, for instance a certification process can be an identification procedure, the participation of a training or passing a driving test. Machines and sensors might gain certification after inspection and maintenance routines. Although in the depiction above the certification is represented with a single arrow, in practice the certification procedure might be more tedious with arrows leading in a bidirectional manner.

After deciding that the certification passes, the issuing organization signs the certificate, that is directly connected to the identity, and creates a cryptographic hash from it. This hash is secured on the distributed ledger with additional meta information, as for instance expiration dates and validity status. In this way, the issuing party is able to revoke a certificate. It is also possible to model trainings, that must be renewed regularly, as for instance fire safety exercises. The process is finished, when the user receives her signed certificate. Our design only approves of storing cryptographic hashes and meta information, that cannot be traced back to a specific person or identity, on the distributed ledger. In particular, the signed certificates are not stored on the distributed ledger, since this opposes the demand of sovereignty.

At this point, one might argue, that the smart contract in our exemplary use case is not able to automatically verify the suitability of an identity, if this information is not stored in the distributed ledger. This is correct. The characteristic transparency of the distributed ledger technology, that is demanded and desired in many applications has a flipside when dealing with personal information. Also, the encryption of the personal data is no solution, since the smart contract program code is accessible, as well and so is the

corresponding decryption key. Therefore, in our approach, we rely on a combination of an on-chain smart contract and an off-chain smart contract (confirm Figure 2). Thereby, a private key to decrypt data is only accessible to the off-chain smart contract, while the on-chain smart contract serves the public key to requesting parties. For evaluation and decryption of information the on-chain smart contract relies on the off-chain part.

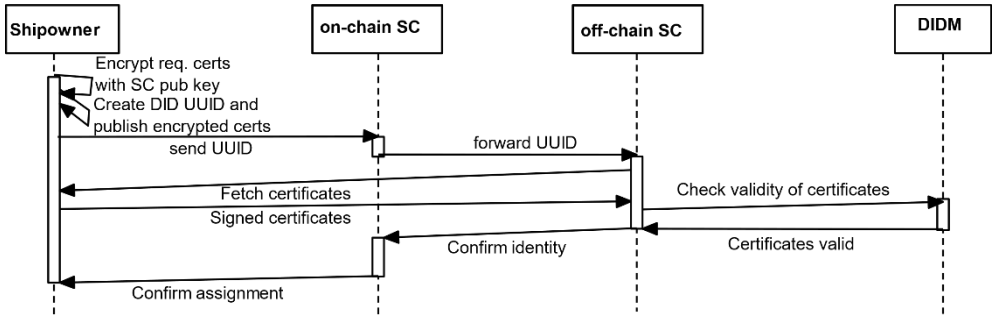


Fig. 2. Application process to a transportation order.

Off-chain and on-chain part of the smart contract are instantiated and operated by the party that offered the transportation order.

The application to a transportation order works as depicted in the sequence diagram of Figure 2: The inland navigation operator generates verifiable representations of her certificates (including the certificates of the ship and other equipment) required for the transportation order. The required certificates are listed in the smart contract. The operator will encrypt her verifiable representations and publish it to some arbitrary cloud storage, as for instance an Amazon bucket. In the following, she generates a DID-UUID, that refers to the stored representations and sends this DID to the on-chain smart contract. The on-chain smart contract activates its off-chain counterpart, that fetches the information from the cloud space and verifies the correctness. It then, confirms the identity to the on-chain smart contract. Thereby the transportation order transits from the application phase into the processing phase.

On first sight, this approach might be overly complex. So, let us discuss the reasons for this exact modeling: Basically, we have two reasons: For once, we aim at thorough sovereignty of personal data and second, we try to protect the identity of parties in the publicly accessible distributed ledger. Let us presume, the inland navigation operator sends her certificate data directly to the on-chain smart contract. The personal information would be practically available forever and a due to the characteristics of the distributed ledger a deletion is not possible. Additionally, as explained before, an optional encryption is useless, since the decryption key would be part of the program code of the smart contract, and thus, accessible by everyone. This can have substantial consequences to the business of the inland navigation operator. Others could mine the information and connect transportation orders directly to a specific identity. Competitors gain deep insights into sensitive business statistics and can use this information to build a competitive advantage. The DID-UUID as introduced in the sequence diagram must not comprise of any personal information and thus, protects the parties against such attacks. With the UUID a person can fetch the information from the cloud storage. However, one can use different cloud storage providers to obfuscate the ownership of certificates,

even more. The certificates are encrypted and can only be decrypted with the corresponding private key of the smart contract. So, no competitor is able to extract the certificate or identity information from the data stored in the cloud space.

These are the reasons for our variant of the DID-UUID and the splitting of the smart contract into an off-chain and an on-chain part. In contrast to our approach the standard in [3] uses the verifiable data registry and thus, the distributed ledger to persist the relation between a public key and a DID. The advantage is that this approach documents the identities to comprehend interactions later on. However, it opens the field for mining attacks as described above.

Often there are statutory and fiscal requirements that demand storage of identity information. Since, the distributed ledger is the immutable storage container that might provide trustworthy documentation to officials, it would be beneficial to document the identity information within the smart contract running on the ledger. We make great efforts to decouple the personal information from the data stored in the smart contract to prevent aforementioned attacks. To mitigate the requirement of documenting the identity of involved parties, the different parties can store the certificates of the other parties locally and secure them by writing a hash into the smart contract. So, for instance, as an additional step after verifying the identity of a party, the off-chain smart contract stores the verified credentials for the period of the retention obligation in a database and secures the content of the database as a cryptographic hash in the distributed ledger. If now, liability issues or tax auditions take place, the auditors only need to request the information from an involved party and can test the authenticity by relying on the distributed ledger.

The sequence diagram in Figure 3 depicts a part of the processing phase of the transportation order. The smart contract encoding the basic conditions of the transportation order leads actively through the transportation process. Depending on the transported freight and additional situation specific conditions, different variants of smart contracts are required. So, for instance we surely need different smart contract to transport bulk good vs piece goods or hazardous goods. However, a templating of smart contract patterns might be feasible. The creator of a transportation order can use programmatic assistance in parametrizing different smart contract variants.

The interesting parts of the sequence diagram are the different situations in which we test or assess parameters and information. As for instance, after the barge arrives at the port, depending on the transported good the freight room is inspected. Trained personal assess the cleanliness of the cargo holds and thus determine its suitability to transport sensitive goods as for instance grain. This results into a report called FRI/CLI report. The report is signed by the inland navigation operator, as well as, from the inspector.

Similarly, it works for the measuring of water displacement to determine the weight of transported freight. This report is again signed by the inland navigation operator and the gauger. In future, there might be calibrated sensors that automatically measure the water displacement. However, also in that case we need the signature of the two parties, since they need to negotiate a consensus, whether the freight room is empty when taking the empty gauge and analogously, when gauging after the ship is completely loaded.

Instead of adding the detailed reports to the distributed ledger a better solution is to store the reports into a database, that is secured with cryptographic hashes written into the distributed ledger. However, if the general test result is submitted to the smart contract, one can implement the process logic in the smart contract and thus react automatically in case for instance the cleaning state of the freight state does not hold up to the requirements of the transported goods.

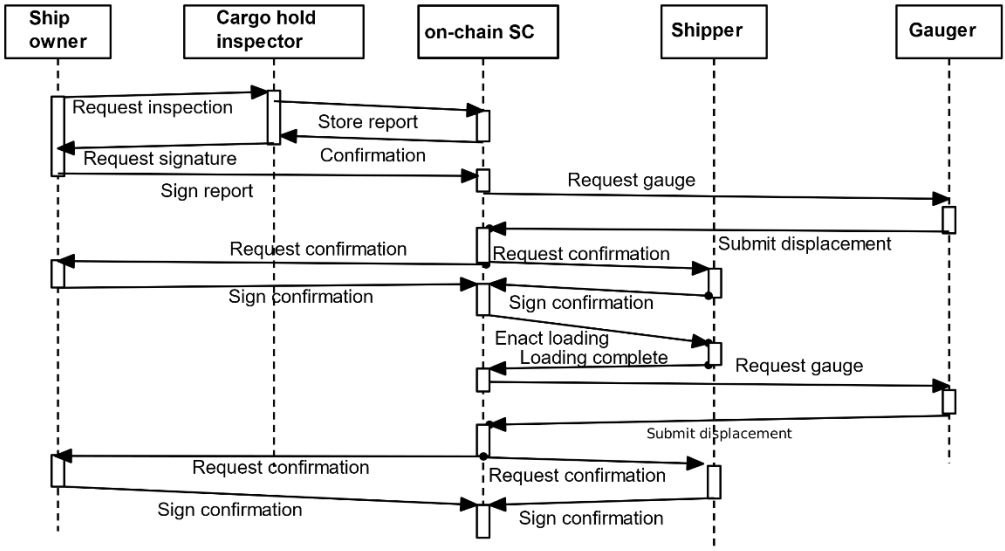


Fig. 3. Transportation process.

Although, we only explained the loading procedure in the sequence diagram of Figure 3 the analogous procedure for unloading has similar requirements on determining the identity of involved parties, that test and assess information.

In conclusion, we can derive the following core features of a decentral identity management in inland waterway transportation from the use case considered:

- The verification and authenticity of identities is of crucial importance in the settlement of inland waterway logistics, since the correctness of information strongly depends on its origins.
- Besides persons and organizations, also machines and sensors, as for instance ships, freight rooms or water displacement sensors need to be equipped with verifiable identities. Only then, one can be sure that a displacement sensor is regularly calibrated and thus, the measurement is trustworthy.
- Smart contracts can be utilized to secure and automate transportation processes. In particular regular processes that follow a fixed pattern.
- The use of decentralized identity management and DID-UUIDs enables a secure authentication while protecting personal and business information.
- To protect personal information, we need to counteract the transparency inherent to DLT. We introduced the idea of using a combination of on-chain and off-chain smart contracts.

One additional advantage of an identity management system that assigns roles to identities, is that it can be used to define data access policies. In the considered transportation process, it is not strictly required that every stakeholder has complete information about the whole process. So, for instance, the inland navigation operator does not need to know the name of the final customer. Instead she only needs to know

the destination port, that than needs to know the final receiver of the freight. In our approach, we use attribute-based encryption to achieve a role specific view on information. Every role is represented by an attribute and the information is separately encrypted. Of course, due to the aforementioned characteristics of the distributed ledger technology, we need to be sensible about the data actually stored on the ledger. Instead, we use distributed databases to store the data and use the decentral identity management and the distributed ledger to handle the assignment of attributes and thus roles to participants.

4. Feasibility of the Approach

There exists numerous challenges, when implementing the discussed approach. In the context of a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis, we can derive the following aspects. The strengths of the approach are the secure and transparent execution of the transportation process and the promotion of digitalization in an environment, that for the time being, works primarily paper-based. The utilization of the distributed ledger technology ensures the transparency and provides also opportunities with regard to process automation and the inclusion of sensors and services. The decentral identity management ensures the participation of arbitrary parties without introducing central authorities. Thereby, the identity and personal (respectively person-related) information are protected from third parties. This represents a strength, since it increases the reliability and reduces the negotiation efforts that are necessary decide on governance and operational details of a central identity management. Regarding the question of opportunities, decentral identity management has the potential to easily provide identities to individuals and organizations, as well as, machines and sensors. Thereby, the utilization of DLT ensures the integrity of the identities.

However, there exists also a number of weaknesses. The most important weakness represents the question, whether the process participants actually adopt the proposed approach. Than, there is the problem of securely mapping an identity to a concrete individual, organization or machine/service. Also, potential scalability issues need to be considered. Right now, the scalability of DLT is an issue. Often the transaction throughput is limited and the required amount of memory space accumulated over a specific operation period is quite substantial. Finally, legal regulations need to be adapted. So, for instance the water police must be entitled to accept cargo documents in digital form and the gauging commission must be extended to cover sensors, as well.

The threats regarding the approach are primarily those, that threaten the distributed ledger technology and cryptographic methods (signature schemes, PKI…) that are utilized in the identity management. Thereby, we can distinguish four classes of risks: system risks, where the ecosystem of the DLT changes. For instance, in case of a proof-of-work consensus a majority gains the computational supremacy. Infrastructure risks, as for instance exploitable bugs in a DLT node implementation. Cryptologic risks, where the cryptography fails, due to bugs in the implementations or due to a disrupting change in the available technology (e.g. post-quantum era). Finally, there exists the user risks, where users jeopardize the security of the DLT system. Besides, the irresponsible handling of public/private key pairs, another example are wrongly implemented smart contract. What if a smart contract runs into a deadlock or is exploitable by third parties? Thus, a formal verification of involved smart contracts and the communication protocol is

important. We proposed an approach to verify the correctness of Ethereum smart contracts utilizing model checking in [6].

5. Conclusion

Often when discussing the beneficial characteristics of the distributed ledger technology and how it enables new use cases and business applications, there is always the claim of a trustworthy network with secure and correct information available to every participant. However, we argue that the correctness and trust in information not only depend on a technology capable of securely storing information, but also on the origin and authenticity of the information. This directly relates to an effective and reliable identity management, that does not counteract the decentral idea of the distributed ledger technology.

The implementation of a combination of on-chain and off-chain smart contracts has the disadvantages of reducing the transparency in the system. As such, the assessment of the identity cannot be comprehended by the inland navigation operator. A potential error in the validation routine or a flawed algorithm will not be noticed, as efficiently, as it is the case in transparent execution. On the other hand, this approach secures the privacy and personal data, as well as, sensitive business information, of the involved parties. The price for the missing transparency can be justified under these conditions. Generally speaking, the acceptance of this price is solely a decision of the involved parties. In this paper, we argued for the need of considering a distributed ledger application in the broader context of sustainability and network cooperation. This includes the field of identity management and role-based encryption to implement transparency where needed and obscurity where necessary.

Acknowledgments

This research is partly supported by project SINLOG, which is funded by the mFund Programme of the German Ministry for Transport and Digital Infrastructure (Project id: 19F2099C), and the b-it foundation[1].

References

[1] Kim Cameron, 'The Laws of Identity'. Microsoft Cooperation, 05-Nov-2005

[2] Dunphy, Paul & Petitcolas, Fabien. (2018). A First Look at Identity Management Schemes on the Blockchain. IEEE Security & Privacy. 16. 10.1109/MSP.2018.3111247.

[3] Sporny, Manu & Longley, Dave & Chadwick, David. (2019). Verifiable Credentials Data Model 1.0. https://www.w3.org/TR/vc-data-model

[4] Reed, Drummond & Sporny, Manu & Longley, Dave & Allen, Christopher & Grant, Ryan & Sabadello, Markus & Holt, Jonathan. (2020). Decentralized Identifiers (DIDs) v1.0. https://www.w3.org/TR/did-core

[5] Panait, Andreea & Olimid, Ruxandra & Stefanescu, Alin. (2020). Analysis of uPort Open, an identity management blockchain-based solution.

[3] http://www.b-it-center.de/

[6] Osterland, Thomas & Rose, Thomas. (2020). Model checking smart contracts for Ethereum. Pervasive and Mobile Computing. 63. 101129. 10.1016/j.pmcj.2020.101129.

Market-related opportunities and challenges for a digital platform model aiming at sustainable execution of last-mile logistics - A use case of B2C deliveries in Germany and Vietnam

Thanh Ha Mai[1], Carsten Dorn[2], Benjamin Wagner vom Berg[3]

1. Introduction

Most of the goods recipients or end consumers are living nowadays in urban areas. The trend of urbanization is evident in the fact that more than 50 percent of the world's population living in cities. This urbanization level will increase to 69 percent in 2050, while this indicator of Europe and the United States will be 85% and 91% respectively [1].

Thus, intensification of transport activities, in particular the inner-city freight traffic is inevitable. Besides urbanization, environmental problems are imposing more pressure upon urban logistics. With its high volume of fossil-fueled vehicles, urban freight transport has a significant negative impact on the environment and daily life of city dwellers through, e.g., CO2 emissions, dust, noise, traffic congestion. Urban goods transport is in particular characterized by last-mile deliveries of parcels to a large number of private households as a result of flourishing electronic commerce.

Last-mile as the final stage of distribution logistics are seen as the most expensive, least efficient and problematic part in the entire logistics value chain [2]. As the main executor of last-mile deliveries, CEP (Courier, Express, and Parcel) service providers are facing challenges in many respects: sophisticated customer requirements, increasing cost and competitive pressure, shortage of drivers and not least stricter environmental regulations.

Therefore, it is high time to work out innovative logistics concepts that support more sustainable development in ecological, economical and social dimensions. A possible solution is a central logistics platform for handling the entire last-mile activities in a city or region. The platform brings main stakeholders of urban logistics (CEP service providers, regional logistics companies, local retailers, municipality, private customers) together, realizing positive effects through transport bundling, the transport system of electric delivery vehicles in combination with hub structure and sustainable crowd-logistics approach [3].

Under consideration of an international application, this research intends to embed the proposed logistics platform model into two different markets. Germany represents an example of a developed European market. In contrast, Vietnam is an evolving market in Asia, which has undergone in recent years dynamic growth in last-mile delivery services through the e-commerce boom. These two markets were chosen for practical reasons. Authors are working in Germany (market proximity) and in particularly, one

[1] Hochschule Bremerhaven, Bremerhaven Germany, tmai@studenten.hs-bremerhaven.de
[2] Hochschule Bremerhaven, Bremerhaven Germany, cdorn@hs-bremerhaven.de
[3] Hochschule Bremerhaven, Bremerhaven Germany, benjamin.wagnervomberg@hs-bremerhaven.de

author is Vietnamese native speaker. Therefore, it is advantageous for the research to search for Vietnamese information source and to conduct telephone interviews with local businesses.

Specific market requirements and market opportunities will be identified, which enables insights for the implementation of sustainable inner-city logistics concepts for the European and Asian markets to a certain extent. The research concentrates on the use case of Business-to-Consumer (B2C) deliveries and conducts interviews with representatives of logistics companies, especially CEP service providers. Their estimations of the market and of the business model are essential for developing an economically viable logistics service platform that keeps up with market situation.

2. State of the Art

2.1. Classical concepts and problems

In the context of the CEP market, last-mile logistics means the transport of shipments from the depot in the target region to the recipient or end customer in the urban areas [5]. Diesel transporters and small trucks up to 3.5 tons are the most established delivery vehicles in practice [6]. Each CEP service provider uses its supply network, so that synergy effects between market players are not exhausted. Multiple deliveries to the same address and insufficient capacity utilization are the consequences [3].

Another common practice is the subcontractor model. It means that some large companies do not offer CEP services from a single source, but partially or entirely outsource delivery on the last mile to subcontractors. It is supposed that the parcel service providers GLS, DPD and Hermes commission their subcontractors to process over 90 percent of their shipment volume. In comparison, these shares are estimated at 5 percent at DHL [7]. Despite benefits as cost-savings, the subcontractor model holds many risks such as loss in quality, uncontrolled adverse effects (underpaid jobs, false self-employment). [8] To the best knowledge of authors, there is so far no example of collaboration be-tween CEP service providers in the last-mile delivery. Logistics service providers compete and there is a lack of attractive incentives for cooperation.

2.2. Innovative solutions

The concept of establishing a new distribution stage (hub or micro hub) before delivery to end customers can be combined with electric vehicles and cargo bikes in urban logistics, to avoid traffic bottlenecks and deliver with zero local emission [10]. For instance, the pilot project of UPS in Hamburg, in which containers are used as a mobile hub and the parcels are delivered from there by sack trucks, was also implemented at other locations [11].

In line with the trend towards digitization, the combined use of novel technologies (e.g. Big Data, Cloud Computing, Artificial Intelligence) enable alternative efficient delivery routes (e.g. in case of traffic problems) and can improve the productivity of deliveries by between 20 and 40 percent [12]. Crowd-logistics is an emerging concept, which is defined as "the outsourcing of logistics services to a mass of actors, whereby a technical infrastructure supports the coordination." [13] Traditional CEP service providers also see opportunities in the concept: DHL tested the crowdsource delivery platform MyWays in

Stockholm with the involvement of city dwellers for parcel delivery [14]. The use of crowd logistics must be designed with consideration for risks of endangered social security system, because crowd-workers are employed as independent contractors and bear all risks and social expenditure themselves [15]. Logistics platforms are more widespread in long haul transport of freight forwarder (e.g. Uber Freight, Saloodo!) [16]. For B2C logistics, the digital cross-company platform is still in the niche application. One example is the platform Tiramizoo, which coordinates same-day deliveries to end customers for retailers (e.g Media Markt) and assigns orders to the local courier service to carry out [17].

3. Platform model and selection of use case

3.1. Company Rytle – product portfolio and strategic direction

Company Rytle offers a wide range of cargo bike models that enable emission-free transport on the last mile as well as tailored solutions for specific requirements (e.g. high/low shipment weight). The highlight of Rytle's product portfolio is the innovative transport system, consisting of four components: electric freight bicycle (MOVR), standardized roll container (BOX), a swap container (HUB) and the APP (as part of digital infrastructure). Rytle cooperates with most CEP service providers in Germany, for both pilot projects and regular operation.

As preparation for the expansion beyond Europe, Rytle enters into cooperation with business partners in the USA, Canada and Asia [19]. The company positions itself as an IT company (as a provider of "Infrastructure as a Service" (IaaS)) and further as a logistics platform provider. Rytle is pursuing the strategy to make itself known internationally through sales of its sustainable transport concept, using the advantage of this market access for the future platform.

3.2. Platform model and its sustainable approach

As aforementioned, the initial idea is to develop a holistic and sustainable platform solution one for all transport needs on the last mile. Besides providing the necessary hard and software components and the sustainable transport system, Rytle takes over the central role of platform operator, coordinating the transport activities and other supporting processes. Hereinafter, three main actors of the platform are demonstrated.

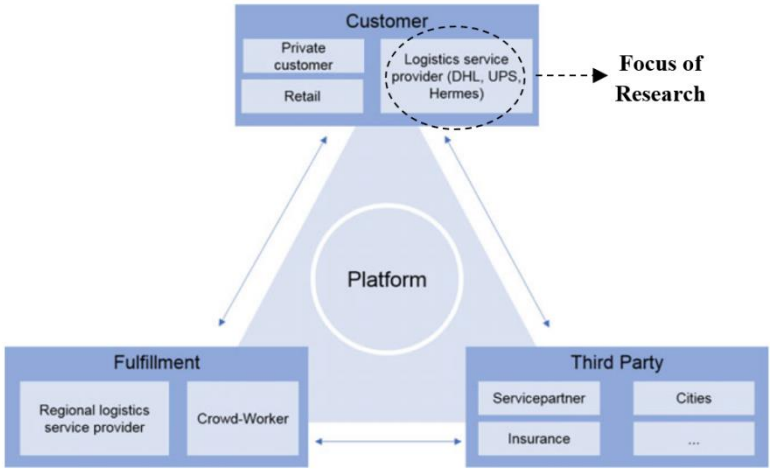


Fig. 1. Main actors of platform model (Wagner vom Berg et al. 2019).

Customers in this model might be predominantly CEP service providers (e.g. DHL, Hermes, GLS) due to their large parcel volume. Research concentrates on figuring out the viewpoint of logistics service providers towards the proposed plat-form model. Other customer groups such as private customers or local retailers should also be mentioned. All of three customer groups can participate in the platform, setting their particular transport requirements. Transport assignments will be referred to the participating regional logistics service provider (Fulfillment) to carry out. Crowd-workers is also part of this group, they should relieve the staffing of local logistics service provider at peak times. The last actor is the third party, which are, e.g. service partners responsible for the maintenance of the vehicle fleet, or also interest groups such as cities (who can favor the operation). Besides B2C deliveries, various other use cases such as rural areas, regional retail, private parcels, were identified [3].

This platform model is a sustainable concept in all three dimensions (ecology, economy and social equity). Firstly, the bundling of different products (e.g. newspapers, parcels) as well as the bundling of services for different clients (local retailers, CEP service providers) can be realized, which results in cost advantages for all stakeholders but also ecological advantages due to the higher utilization of loading capacity and reduction of distances. Ideal scenario, where one singular regional logistics service provider carries out all the transports within a city based on a white label solution, will leverage once again the positive effects of transport bundling. Environmental problems can also be mitigated by the use of cargo bikes and electric driven delivery system. With the approach of sustainable crowd logistics, a competitive but also employee-friendly working model in terms of social sustainability can be realized [20]. Significant efficiency gains on the part of the logistics service provider are also to be expected, since staff shortages can be eliminated and working conditions of existing staff will be improved permanently.

4. Identification of specific market requirements

4.1. Research Methodology

Expert interviews were applied to gain market insights and perspectives of German and Vietnamese logistics service providers about the platform. In preparation for that, a thorough internet search was conducted to identify the major logistics companies in the field. Suitable contact persons were founded and addressed on the social network LinkedIn. In parallel, a standard guideline, which includes questions about opinions of interview partners regarding transport concept and platform mod-el, was created. By answering these questions, the interviewee also provides vital facts about the respective market. Finally, six interviews were conducted from October to December 2019 (Table 1).

Company	Job of interviewed person	Delivery area
Weser Eilboten	Project Manager	Regional (Germany)
Deutsche Post DHL	Senior Expert	Interregional (Germany)
Giao Hang Nhanh (GHN)	Former Co-Founder & CEO	Interregional (Vietnam)
CITIPOST Nordwest	Managing Director	Regional (Germany)
DHL eCommerce Vietnam	Operations Director	Interregional (Vietnam)
Company X* (anonymized)	Operations Director	Interregional (Vietnam)

Table 1. List of interviewed companies (in chronological order).

Through the interviews with representatives of two regional logistics service providers (Weser Eilboten, CITIPOST Nordwest), it can be stated that these companies offer many different logistics services (e.g. freight forwarding, press logistics) and they are already working as a subcontractor for the parcel delivery company. Mainly, CITIPOST Nordwest, by order of GLS, has applied since 2018 the transport sys-tem of Rytle to deliver parcels emission-free in the city areas of Oldenburg. [21] The remaining logistics service providers operate interregional and are important delivery partners of e-commerce. Among all interviewed companies, Giao Hang Nhanh (GHN) is the unique technology-orientated start-up for last-mile delivery services.

4.2. Key findings

Market background and trends

German market: E-commerce in Germany is one of the largest in Europe, with a total sales of 65.1 billion euros (two-thirds related to B2C sales) [22]. Online shops with high turnover as Amazon, Otto, Zalando are important customers of large CEP service providers such as DHL, Hermes, DPD, UPS and GLS. In 2017, CEP service providers generated 19.4 billion euros turnover through the transportation of 3.35 billion CEP consignments (over 50 percent therefrom assigned to B2C shipments) [23]. CEP service providers are facing many ecological, economic and social challenges: declining revenue per package, bad reputation due to precarious working conditions and low wage level through subcontractor model, entry

bans for old diesel vehicles in large cities, etc. As the market leader, Deutsche Post DHL invests in electric van production (StreetScooter), testing the use of cargo bikes and city hubs (cubicycles), in the effort to achieve the goal of zero-emission logistics by 2050 [24]. CEP companies in Germany also recognize the potential to improve their last-mile process through the combined use of cargo bikes with a micro-depot concept (e.g. KoMoDo project) [25]. Another noticeable market strategy is the enlargement of a nationwide network of parcel shops, post offices and packing stations to enable flexible parcel receiving. Nevertheless, CEP service providers in Germany have difficulties in implementing same-day delivery service (which is currently offered by Amazon and many other online retailers) with their existing processes. As a re-sult, a new business model (Tiramizoo, Liefery) emerges to occupy this market segment. In line with the business model of Uber Freight, Amazon launched the "Amazon Flex" platform in Germany, where every private person can register as a delivery partner and earn money by delivering Amazon packages with their own vehicle. There is criticism that Amazon Flex will lead to precarisation of working conditions of delivery personnel [26].

Vietnamese market: In 2018, e-commerce in Vietnam generated B2C sales of 8.06 billion US Dollars. The growth potential has not yet been exhausted, as the number of paying customers of online trade is currently only around 35 percent of the entire population [27]. Similar to many Asian countries, Cash-on-Delivery (COD) is the most used payment method (in 80 percent of online purchases) [28]. This is one of the biggest challenges for e-commerce and last-mile logistics in Vietnam since many customers decide to cancel the purchase at the time of delivery and many risks occur in regards to the handling of cash. Another challenge is the poor road infrastructure for transport activities (both long haul and last-mile transport) [29]. For-merly, large postal and parcel service providers dominate the market of B2C parcel delivery. Vietnam Post (VNPost), the largest state-owned postal company, leads over other market participants, thanks to a nationwide delivery network. However, the company has missed its competitive advantage due to conservative strategies (neglect of investment in new IT technologies) and inefficient process flows. As a new market entrant, start-up GHN discovers the opportunities for itself and manages to offer its delivery services for 63 provinces and cities in Vietnam [30]. A critical market trend is also the effort of logistics companies to shorten delivery times. Finally, the market is conquered by a crowdsourced delivery model, operated by foreign businesses such as Singapore-based technology company Grab [31]. It follows the model of Uber, using a large pool of private persons as crowd deliverers.

Relevant market requirements of German and Vietnamese market in comparison

The main differences between German and Vietnamese last-mile logistics are grouped into two dimensions, namely market structure and required customer service (Figures 2 and 3). Illustrations in the form of polarity profiles demonstrate ex-tent or characteristic (importance, availability) of market requirements in each country. The rating scale is from 1 to 3 (1 is unimportant and/or not available and three is very important and/or with strong presence). The scores base on an independent assessment by the authors.

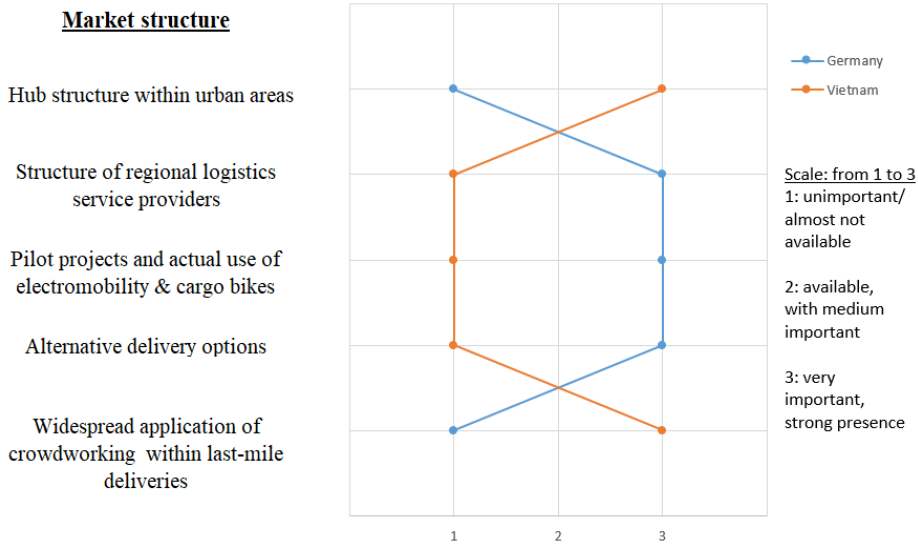


Fig. 2. Differences in market structure between German and Vietnamese market.

Regarding the dimension of market structure: First of all, the hub structure within last-mile delivery is currently pursued but still almost not available in the network of CEP service providers in the German market (score 1). In contrast, hub structure is already standard practice in the Vietnamese market for every CEP service provider, through the application of inner-city delivery stations (score 3). Secondly, the structure of regional logistics service providers is typical in Germany, which matches well the concept of the platform. However, these companies hardly exist in Vietnam due to the tough competition on the market. Thirdly, pilot projects and actual use of electro mobility and cargo bikes are current issues in Germany, whereas the dominant delivery vehicle for last-mile in Vietnam is a motorbike. Alternative delivery vehicles still do not play a role in the company's strategy in the short and medium-term. Alternative delivery options e.g. parcel shop, parcel station are more critical for German customers than fast delivery and hence very widespread [32]. In Vietnam, delivery alternatives are nearly not available, due to low customer requests. The Vietnamese market is more advanced in the application of crowd logistics within last-mile deliveries.

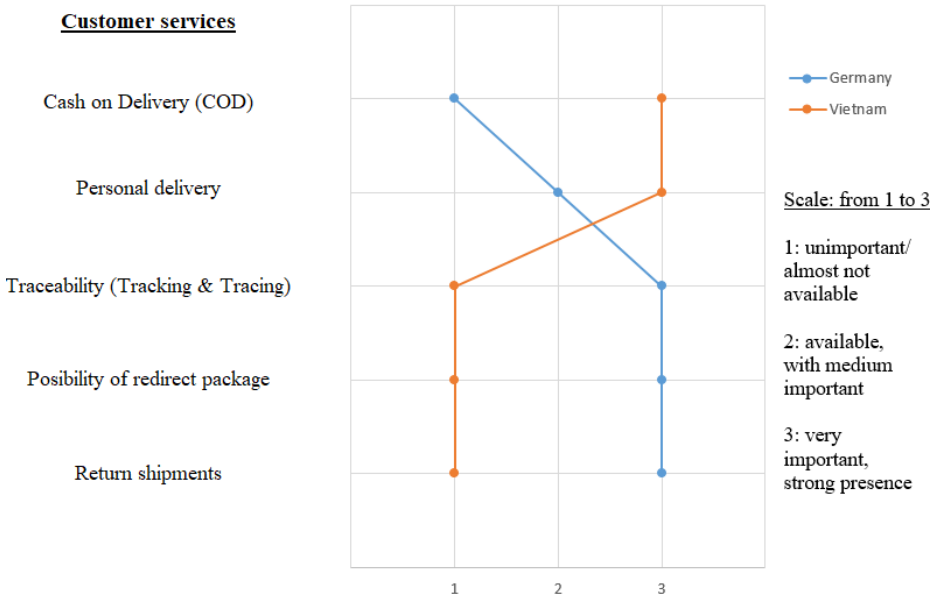


Fig. 3. Differences in customer services between German and Vietnamese market.

Regarding the dimension of customer service: Difference can be seen firstly in the demand of Cash on Delivery (COD), as an extra service at the time of delivery. In Vietnam, COD is the most used payment method in online trading (in 80 percent of online purchases) [28]. As opposed to that, COD is irrelevant in Germany because of other common-used payment methods e.g. Paypal, invoice, direct debit. Consequently, personal delivery is highly vital within last-mile deliveries in Vietnam, whereas it plays medium importance in Germany due to alternative delivery options. Consignment tracking (Traceability) is now a standard service offered by German CEP service providers but almost not available in Vietnam. The same tendency can be seen in other services (possibility of redirect package and return shipments), which are more relevant in the German market, but currently are of minor importance in the Vietnamese market.

Opportunities and challenges for RYTLE platform

There is a consensus among all interviewees that the scenario of one singular regional logistics service provider carrying out all the transports within a city is unrealistic due to the current competitive situation and antitrust reasons. It is more likely for the proposed platform that several regional logistics service providers participate in the platform as transport executors. The last-mile delivery based on white label solution proves to be undesirable and some CEP service providers will contemplate the participation only if brand recognition and improvements in costs and service quality are guaranteed.

In detail, interviewed regional logistics service providers (Weser Eilboten and CITIPOST Nordwest) showed interest and willingness to participate in the platform. Nevertheless, they express concern that large CEP companies still gain attractive profits from last-mile transport and hence, the acquisition of this target group can be challenging. It was confirmed in the interview that the Deutsche Post DHL see last-mile

delivery as their core competences. The representative of the company doubts whether platform model and concepts such as transport bundling and crowd logistics will bring more efficiency actually. High adaption costs for IT interfaces was also mentioned as a significant obstacle for DHL to participate in the platform. Furthermore, interview partners from CITIPOST Nordwest see more market opportunity for the application of a platform for rural areas.

Vietnamese interview partner confirmed that the structure of regional logistics service providers does not exist on their market. Start-up GHN, as a successful company in this sector, also applies a crowd-logistics approach with its subsidiary company to counteract the fluctuating order volume and to achieve synergy effects of two business models [33]. Therefore, the CEO of this start-up does not see profits for his company through the platform. Company X (anonymized by request of interviewee) is an interregional logistics service provider, see its participation instead as transport executor. This indicates that adjustments, regarding the role of actors with-in the platform model, should be considered to adapt to specific market conditions.

5. Conclusion and Outlook

For the implementation of the business model in the German market, Rytle's platform model incorporates promising concepts such as electromobility, hub structure and sustainable crowd logistics approach. The challenge lies in the acquisition of CEP service providers since they still doubt the positive effects and benefits of transport bundling and prefer executing last-mile deliveries themself. There is a fundamental problem for the application of platform in Vietnam, namely different market structures and trends (the dominance of large interregional companies, the existing hub structure in inner urban areas, the widespread of the crowdsourced delivery model). Mainly, it will not be easy for the original approach of sustainable crowd-logistics to succeed in this market.

A limitation of the research is the small sample of qualitative interviews due to time reasons and the low response rate of contacted companies. CEP companies are an important target group for the proposed platform and need to be differentiated according to their business strategies. It appears to be a worthwhile task to iterate the research with a larger sample of companies and additionally involving first results from the pilot phase of NaCl project to raise more interests and confidence in the digital business model.

Further direction of this research can be seen in a broader application field of the platform. Besides coordinating the last-mile transport activities, a holistic and sustainable platform can further take over other important functions such as supporting regional retailers to digitalize their product offering, reinforce regional retail and transport sector. By this way, environmental effects of transport activities will be mitigated, whereas regional value creation and regional supply structure will be promoted. This is the aim of pending research project R3 (Resilient Regional Retail), which is going to start in spring 2021 at the University of Applied Sciences Bremerhaven [34].

Acknowledgements

The paper is part of the NaCl project. NaCl is a joint project of the University of Applied Sciences Bremerhaven (project coordinator), Rytle GmbH and Weser Eilboten. The project is funded by EFRE within the program “Applied Environmental Re-search (AUF)” of the city of Bremen. (City of Bremen, 2018).

References

[1] United Nations Website, https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf, last accessed 2020/07/17.

[2] Gevaers, R., Van de Voorde, E., Vanelslander, T.: Characteristics of innovations in last mile logistics – using best practices, case studies and making the link with green and sustainable logistics. European Transport Conference, Noordwijkerhout (2009).

[3] Wagner vom Berg, B., Oetjen, N., Schopka, K., Hanneken, F., Reiss, N., Hollmann, R.: Platform Sustainable Last-Mile-Logistik – One for ALL (14ALL). In: Schaldach et al.: Advances and New Trends in Environmental Informatics, pp. 67-78. Springer Berlin Heidelberg, Berlin, Heidelberg (2019).

[4] Wagner vom Berg, B., Oetjen, N., Schopka, K., Hanneken, F., Reiss, N., Hollmann, R.: A Sustainable CRM approach to a Crowd Sourced Last Mile Logistics Platform (NaCl). In Proceedings of 33rd EnviroInfo conference 2019, pp. 324-335, Shaker Verlag, Kassel (2019).

[5] Arnold, D., Isermann, H., Kuhn, A., Tempelmeier, H., Furmans, K.: Handbuch Logistik. 3rd edn. Springer, Berlin (2008).

[6] Thaller, C., Telake, M., Clausen, U., Dahmen, B., Leerkamp, B.: KEP-Verkehr in urbanen Räumen. In: Proff, H., Fojcik, T.M. (eds.): Innovative Produkte und Dienstleistungen in der Mobilität. Technische und betriebswirtschaftliche Aspekte, pp. 443-458. Springer Gabler, Wiesbaden (2017).

[7] WIK Consult: Development of Cross-border E-commerce through Parcel Delivery. Study for the European Commission, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs. Final Report. In: https://www.wik.org/fileadmin/Studien/2019/ET0219218ENN_ParcelsStudy_Final.pdf, last accessed 2020/07/17.

[8] Bennühr, S.: Das Ende des Subunternehmermodells. In: https://www.dvz.de/rubriken/mei-nung/detail/news/das-ende-des-subunternehmermodells.html, last accessed 2020/07/17.

[9] Park, H., Park, D., Jeong, I.: An effects analysis of logistics collaboration in last-mile networks for CEP delivery services. Transport Policy Transport Policy, Volume 50, 115-125 (2016).

[10] PwC PricewaterhouseCoopers: Aufbruch auf der letzten Meile. Neue Wege für die städtische Logistik. In: https://www.pwc.de/de/transport-und-logistik/pwc-studie-aufbruch-auf-der-letzten-meile.pdf, last accessed 2020/07/17.

[11] HSBA Hamburg School of Business Administration: Last-Mile-Logistics Hamburg – In-nerstädtische Zustelllogistik. In: https://www.hsba.de/fileadmin/user_upload/bereiche/for-schung/Forschungsprojekte/Abschlussbericht_Last_Mile_Logistics.pdf, last accessed 2020/07/17.

[12] DHL Homepage, https://www.dhl.com/global-en/home/insights-and-innovation/insights/shortening-the-last-mile.html, last accessed 2020/07/17.

[13] Mehmann, J., Frehe, V., Teuteberg, F.: Crowd-logistics – A Literature Review and a Maturity Model. In: Kersten, W., Blecker, T., Ringle, C.M. (eds.) Innovations and Strategies for Logistics and Supply Chains, pp. 117–145. Epubli GmbH, Hamburg (2015).

[14] DHL Homepage. https://www.dhl.com/en/press/releases/releases_2013/logistics/dhl_crowd_sources_deliveries_in_stockholm_with_myways.html#.Xnze9XdFw2y, last accessed 2020/07/17.

[15] Eichhorst, W., Hinte, H., Rinne, U., Tobsch, V.: Digitalisierung und Arbeitsmarkt: Aktuelle Entwicklungen und sozialpolitische Herausforderungen. IZA Standpunkte Nr. 85 (2016).

[16] Saloodo! Homepage, https://www.saloodo.com/de/presse/saloodo-stellt-mitglieder-des-neu-geschaffenen-advisory-boards-vor-2017-06-19/, last accessed 2020/07/17.

[17] Tiramizoo Website, https://www.tiramizoo.com/en/, last accessed 2020/07/17.

[18] Rytle GmbH Website, https://rytle.de/, last accessed 2020/07/17.

[19] German Accelerator Website, https://www.germanaccelerator.com/blog/pitch-perfect-30-german-startups-presented-at-our-german-startup-night-asia-edition/, last accessed 2020/07/17.

[20] Wagner vom Berg, B., Moradi, M.: Sustainable labor conditions by Gig-economy – Case Study: Sustainable Crowdlogistics (NaCl). Weizenbaum Conference, Bremerhaven (2019).

[21] GLS Website, https://www.gls-newsroom.de/de/news/gls-startet-emissionsfreie-zustellung-in-oldenburg/s/63099f42-14fe-4e36-a1d5-bc3b967228bc, last accessed 2020/07/17.

[22] Statista Website, https://de.statista.com/outlook/243/137/ecommerce/deutschland#market-revenue, last accessed 2020/07/17.

[23] BIEK Bundesverbandes Paket und Expresslogistik Website, https://www.biek.de/publikationen/studien.html, last accessed 2020/07/17.

[24] Deutsche Post DHL Website, https://www.dpdhl.com/en/media-relations/media-center/videos/our-mission-2050-zero-emission.html, last accessed 2020/07/17.

[25] KoMoDo Homepage, https://www.komodo.berlin/, last accessed 2020/07/17.

[26] Eurotransport website, https://www.eurotransport.de/artikel/amazon-flex-in-muenchen-paketzustellung-als-hobby-10904134.html, last accessed 2020/07/17.

[27] Vinalink Logistics Homepage, https://www.vinalinklogistics.com/ban-tin/bao-cao-logistics-viet-nam-2018-logistics-va-thuong-mai-dien-tu-1743.html, last accessed 2020/07/17.

[28] European Chamber of Commerce in Vietnam – Eurocham Website, https://www.eurochamvn.org/node/17882, last accessed 2020/07/17.

[29] Brands Vietnam Website, https://www.brandsvietnam.com/congdong/topic/14957-Top-nhung-kho-khan-cua-nganh-ban-le-truc-tuyen-nam-2018, last accessed 2020/07/17.

[30] GHN Homepage, https://ghn.vn/pages/gioi-thieu-ve-ghn, last accessed 2020/07/17.

[31] Grab Homepage, https://www.grab.com/sg/, last accessed 2020/07/17.

[32] PwC PricewaterhouseCoopers: Wege aus dem Paketdilema. In: https://www.pwc.de/de/transport-und-logistik/pwc-paketpreisstudie-2018.pdf, last accessed 2020/07/17.

[33] AhaMove Homepage, https://ahamove.com/, last accessed 2020/07/17.

[34] Hochschule Bremerhaven, https://www.hs-bremerhaven.de/hochschule/aktuelles/news/2020/forschungsprojekt-soll-regionalen-einzelhandel-staerken/, last accessed 2020/09/10.

Visualization of greenhouse gas emissions for the means of transport airplane, car, train and coach by use of accessibility graphs.

Malte Christiansen[1], Jochen Wittmann[2]

1. Introduction

Business trips by university members are part of everyday working life. Attending international conferences is indispensable for scientific exchange. In addition, there are projects with the corresponding commitments to business trips on the occasion of meetings and presentations. This also applies to employees at the HTW Berlin (University of Applied Sciences). They undertake business trips all over the world. These journeys are often and as a matter of course made by plane. According to the environmental guidelines of HTW Berlin, public transportation should be preferred, but only if it is economically justifiable.[3] [2].

A research group at the HTW Berlin collected the travel data for 2017 and illustrated the distribution of means of transport [3]. Their raw data were kindly made available to this project and form the starting point for this project.

When comparing different means of transport, arguments are usually based on time and costs. For example, the probably most frequently used route planner of Google only offers exactly these two target criteria for a route optimization. Certainly, time and distance are two important and comprehensible arguments, but the comparison could be extended to include the effects on climate and health in order to make the ecological dimension transparent when choosing a means of transport.

In order to provide employees who travel by alternative means of transport with an argumentation aid, this project wants to visualize the difference between planes and alternatives in terms of greenhouse gas emissions by maps. The car, coach and train were chosen as alternative means of transport. These are common means of transport.

The aim is to integrate this information into the process of planning and booking business trips in order to raise awareness of the problem of CO_2-saving mobility among those involved.

In the following sections, the visualization process of the accessibility graphs themselves will be explained briefly. This is followed by the procedure for data selection and integration as well as the possibilities and problems of using the corresponding tools under ArcMap. The boundary condition of the study is also discussed, especially the limitation to open-source data material. First results and a discussion of the difficulties in the implementation concludes the paper.

[1] Hochschule für Technik und Wirtschaft Berlin, Degree course Umweltinformatik, Wilhelminenhofstraße 75A, 12459 Berlin, Germany, s0563114@htw-berlin.de

[2] Hochschule für Technik und Wirtschaft Berlin, Degree course Umweltinformatik, Wilhelminenhofstraße 75A, 12459 Berlin, Germany, wittmann@htw-berlin.de

[3] This was the case at the time the project was implemented (first half of 2019). As of 01.01.2020, "Goal[s] that can be reached by train within six hours travel time will no longer be undertaken by air. For all air travel that continues to be undertaken, greenhouse gas emissions will be offset by investment in climate protection projects" [1].

2. Method

Accessibility graphs serve as the basis for the visualization method used. Accessibility graphs are used in cartography to show the distance that can be covered from a given point. See Figure 1 for an example, which shows the oldest known representation of an accessibility graph. It shows the travel time in different regions of the world. The starting point is London. From this point it is calculated how much time is needed on the given route network. In a certain way the representation is based on each other. The dark green area is reached fastest, followed by a light yellow ring and so on [4].

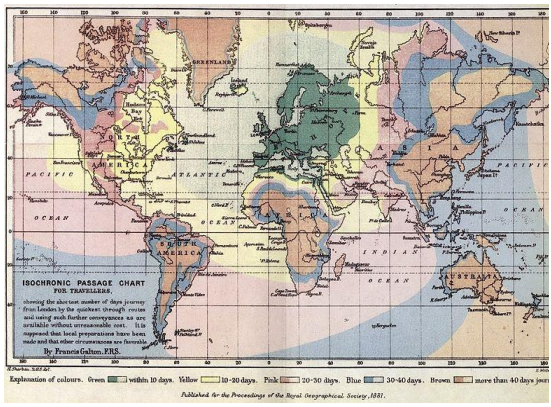


Fig. 1. The first known accessibility map from 1881, showing the travel time to different regions of the world [4].

This technique is still used today to visualize which places can be reached within a given time. In this thesis the attribute "time" is not given as a parameter for determining the extent of the accessibility zones. Rather, an upper limit for the distance to be travelled is calculated first. Afterwards this value is used to visualize how far one can get on the given road network while keeping this distance maximum. The following section explains the procedure in detail.

3. Procedure

After illustrating the method, the description of the procedure follows. First, a route network was required. Secondly, the limit value for the corresponding accessibility graphs had to be calculated. In the third and last step the accessibility graphs could be created.

3.1. Create base layer

An accessibility graph calculates the possible route using a given route network. This project compares the flight route with the means of transport car, coach and train, so that two route networks are needed. One is a road network for cars and coaches and the other a rail network for trains.

The basis is data from the OpenStreetMap project, which was obtained from the Geofabrik download servers [5]. The European countries and regions were downloaded as shape files. From these shape-files the line features needed for the project were selected. The wiki pages of the OpenStreetMap project served as orientation.

- For the road network the Key "highway" was used with the following values:
  - motorway
  - motorway_link
  - trunk

- trunk_link
- primary
- primary_link
- secondary
- secondary_link
- tertiary
- tertiary_link
- unclassified

- For the rail network the Key "railway" was used with the following values:
  - rail

The aim was to deliberately include low-ranking traffic routes, but to reduce the amount of data to a manageable size. The selection of the attribute values listed here is certainly practicable for this prototype in terms of a feasibility study. However, for integration into the target system, which is to cover the entire business trip planning process, it must certainly be checked whether the quantity of necessary values cannot be further reduced without causing significant reductions in the accuracy of the result map. For example, it might make sense to scale the underlying route network according to the desired scale of the target map: In the case of a long journey, the smallest route connections will certainly be less relevant for the final result than for a shorter distance.

After the amount of data had shrunk considerably, the countries and regions could be combined into a map and stored in a geodatabase to facilitate further processing.

3.2. Calculation of the CO_2 factors

A limit value was calculated for each selected means of transport. This value is calculated from the initial value and the respective consumption per passenger kilometre. The initial value is defined as the quantity of greenhouse gases emitted for a flight from Berlin to Munich. This was the most frequently used connection of HTW employees in 2017 [3]. The flight distance from Berlin to Munich is 528 km. The corresponding amount of greenhouse gases per passenger kilometre is taken from Table 1 and offset against each other.

$$528\,km * 201\,g/Pkm = 106128\,g/P$$

One person emits 10.6 kg of greenhouse gases on a flight from Berlin to Munich.

The following formula has now been used to calculate the corresponding distances for the other three means of transport:

$$\frac{initial\ value}{greenhouse\ gases} = distance$$

As example for the train:

$$\frac{106128}{36} = 2948$$

Table 1 contains the correspondingly calculated distances entered as limit values for the respective means of transport. The basis for this was the now generally accepted average emissions as compiled by the Federal Environment Agency [6].

Means of transport	Greenhouse gases (g/Pkm)	Workload	Distance
Airplane	201	82 %	528 km
Car	139	1.5 persons per car	763.51 km
Coach	32	60 %	3316.5 km
Train (long-distance traffic)	36	56 %	2948 km

Table 1. Average emissions of individual means of transport in 2017 [6].

3.3. Creating distance graphs

After the calculation of the limit values and the successful creation of the route network, the distance graphs could be started. The extension "ArcGis Network Analyst" from ArcGis was used for this purpose [7].

The "ArcGis Network Analyst" requires a network dataset consisting of connected edges (lines) and connection nodes (points). In this project, the road network and the rail network are converted into a network dataset in order to perform a network analysis.

During the analysis time factors and distances can be specified. This project works with given distances. The value in the column Distance in Table 1 is taken as distance for the respective means of transport.

Due to the considerable amount of data, it was not possible to create polygons for the coach and the train. These were subsequently created from the line nets using the "Feature in Polygon" tool.

4. Results

Figures 2, 3 and 4 show the respective accessibility graphs for the means of transport car, coach and train. In particular, Figures 2 and 3 show that the possible distance travelled is greatly increased if the coach or train is selected as the vehicle. By contrast, the difference between car and train can be considered as not very large. Instead of travelling from Berlin to Munich by plane, a car could travel from Berlin to Innsbruck (see Figure 2).

Fig. 2. Accessibility graph for a passenger car if 10.6 kg greenhouse gases are available.

Fig. 3. Accessibility graph for a coach if 10.6 kg greenhouse gases are available.

Fig. 4. Accessibility graph for a train if 10.6 kg greenhouse gases are available.

5. Discussion

During the project there were a number of factors that influenced the quality of the visualization more or less. On the one hand, there was the already mentioned limitation in computing power, on the other hand, there was the source material of the road network, and finally, there were also the flat-rate greenhouse gas values used. This will be discussed in this section.

5.1. Computing power

The computers in use reached the limits of their calculation capabilities for the "ArcGIS Network Analyst". It was not possible to calculate a polygon for the train and the coach in a reasonable time. The calculation of an accessibility graph was aborted on a standard PC after about 36 hours of computing time. Instead, a detour had to be made through the calculated lines, which is a potential source of error, as some areas are not displayed correctly with this procedure. This concerns for example areas in Norway and Sweden. However, this is not a major limitation with regard to the accuracy requirements of the prototype.

For example, the polygon is only used for visualization, but does not influence the actual calculation of accessibility. In reality, there are no roads or rails in the spaces in between, which would allow someone to enter these spaces. The polygons therefore only enlarge the visual impression of the accessible area.

5.2. Source material

The source material is based on countries that Geofabrik assigns to the European area. So Georgia, Russia (the non-European part) and Turkey became part of the raw material. The borderline of the project is not very clear. But why are countries like Azerbaijan or Kazakhstan no longer part of the source material? The question remains unanswered on the pages of the provider [5]. Since the selection of the countries was quite arbitrary. For subsequent projects the approximate dimensions should be estimated in advance. This should serve as a guideline, which countries should be considered to create a base layer.

A striking difference between Figures 3 and 4 is that there is a connection between France and the UK for the rail network, but not for the road network. This means that the coach map does not show the theoretically possible route to the UK. This is because the coach could only cross the tunnel by train or cross the Channel by ferry. The coach itself must use other means of transport. Since no roads cross the Channel, the Network Analyst cannot easily take this route into account and could have used temporary connections.

5.3. Greenhouse gases for each mode of transport

The values chosen are based on average values from the Federal Environment Agency. The actual amount of greenhouse gases emitted for a flight between Berlin and Munich could be higher. More greenhouse gases tend to be emitted during take-off and landing than during the flight itself. This means that short-haul flights (less than 750 km) emit more greenhouse gases on average than long-haul flights. The Federal Environment Agency's table, however, only gives a value for aircraft as a means of transport. [6] When applying this to all business trips, it is therefore essential to differentiate according to the length of the flight or a more detailed classification of the flight. For the feasibility study of the prototype presented here, this specification was initially omitted, especially since the trips in the database were not queried as the connection flown was a direct flight with or without a stopover.

Furthermore, the load factor of a passenger car on an actual business trip can be higher than the 1.5 persons per passenger car indicated here. The range of the vehicle increases, the more people are sitting in the vehicle. However, this is information that is available in the use case under consideration when planning a concrete business trip and can be queried by the user as an additional parameter. In this way, a CO_2 value specific to the business trip currently under investigation and thus a specific potential range can be easily determined.

6. Conclusion and outlook

Despite the above-mentioned limitations, the difference in the possible distance between air travel and other means of transport, especially coach and train, is clearly visible. With the amount of greenhouse gases emitted by a flight from Berlin to Munich, the train would reach as far as Lisbon and the coach as far behind Moscow into the interior of Russia (e.g. Ufa).

Initial presentations at the university have shown that this type of visualization certainly promotes awareness of the problem by clearly demonstrating the potential for saving greenhouse gases by avoiding air travel. Facing the climate crisis every company should take action to reduce their greenhouse gas

emission. As shown there is a great potential for a reduction if alternative forms of business travel are considered.

References

[1] Mobilität, HTW Berlin. https://www.htw-berlin.de/einrichtungen/zentrale-hochschulverwaltung/technische-dienste/organisation-atd/umweltmanagement/aktivitaeten-und-tipps/mobilitaet/ (accessed Jul. 09, 2019).

[2] Umweltleitlinien der HTW Berlin, HTW Berlin, Jan. 23, 2017, Accessed: Jan. 26, 2020. [Online]. Available: https://www.htw-berlin.de/fileadmin/HTW/Zentral/ZHV_IIQM_Qualitaetsmanagement/08_Umweltleitlinien_final.pdf.

[3] M. Fronk, A. Güccük, M. Höhne, A. Motzu, and A. Zagorski, Erfassung und Auswertung der mit Dienstreisen verbundenen Umweltauswirkungen der HTW Berlin, Berlin, 2019.

[4] F. Galton, Isochronic Passage Chart, Royal Geographical Society, London, 1881.

[5] Geofabrik Download Server, Geofabrik Download Server. https://download.geofabrik.de/europe.html (accessed Jul. 09, 2019).

[6] Emissionsdaten, Umweltbundesamt, Mar. 13, 2018. https://web.archive.org/web/20190718134549/https://www.umweltbundesamt.de/themen/verkehr-laerm/emissionsdaten (accessed Jul. 09, 2019).

[7] ArcGIS Network Analyst. Esri.

How to consolidate sustainable mobility platforms in rural areas?

Johannes Schering[1], Julian Rawe[2], Ali Akyol[3], Cedrik Theesen[4], Jorge Marx Gómez[5]

1. Introduction: Relevance of sustainable mobility in rural regions

Mobility transition in rural areas is a very tough challenge. Future mobility solutions are mostly considered from the urban perspective but do not lead to a step forward on the countryside. Public transport, bike- or car-sharing are available to reach city centres 24/7. Contrary to the widespread expectation, most of these mobility solutions are not available for the majority of the population. More than 57 million inhabitants of Germany are living in rural areas. Numbers reveal that 90 percent of Germany's surface belongs to rural areas [1]. Furthermore, each household owns at least one car on average [2]. On this account, it is not surprising that there is a high dependency on private cars outside of urban regions. As almost everyone owns an automobile, is there any potential for ride-sharing in the countryside? The recently finished inter- and transdisciplinary research project ***NEMo (Sustainable fulfilment of mobility demands in rural areas)*** [6] was established to find an answer to this question.

The **related work** in this field of research often focuses solely on mobility platforms in the urban context. A publication by Forbes et al. as part of the SUPERHUB project (SUstainable and PERsuasive Human Users moBility in future cities) aims at the development of an open urban mobility hub which incorporates all available mobility offers. Furthermore, gamification and social network approaches are supposed to increase the environmental awareness of the users [3]. Masuch et al. do also propose an open mobility platform for the urban sector. Moreover, an overview of the system architecture and potential use cases of the platform are presented [4]. The approach of Lamberti et al. to an urban multimodal mobility platform is unconventional as distributed ledger technology is used to manage the platform data [5]. Nonetheless, it becomes not clear how to adopt all these solutions to rural areas or how to reach an increasing number of active users by integrating new services that fulfil other demands beyond the improvement of the mobility situation. Therefore, we propose a comprehensive mobility platform which promotes active citizen participation and integrates services besides mobility ones to strengthen local communities.

[1] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, johannes.schering@uni-oldenburg.de
[2] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, julian.rawe@uni-oldenburg.de
[3] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, ali.akyol@uni-oldenburg.de
[4] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, cedrik.theesen@uni-oldenburg.de
[5] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, jorge.marx.gomez@uni-oldenburg.de
[6] https://nemo-mobilitaet.de

The applied **research method** in this paper is based on the design science research approach according to Hevner et al. [6]. Derived from that approach Peffers et al. developed a design science research methodology which consists of six main working steps. In the first step the motivation and the potential problems are described. The second step explains the goals of a feasible solution. Third, an artifact is outlined and generated. The fourth step demonstrates the capabilities of the created artifact. Fifth, the artifact is evaluated. Finally, the artifact is communicated to the scientific audience in step six. [7] The concept for a comprehensive mobility platform in this paper follows the first three steps and the last one. It has to be mentioned that we only perform the first part of step three what includes the outline but not the development of an artifact. The remaining steps three (development), four and five (demonstration and evaluation) of the research methodology explained above will be part of future research.

1.1. Potential of ride-sharing solutions in rural areas: NEMo project

Except for the private car, there are nearly no appropriate mobility solutions available in rural areas. Public transport is mainly focused on school traffic, distances between single bus and train stops are too long and the frequency is not sufficient. Especially at weekends or in evening hours, the private car is the only option. [8] It might well be that ride-sharing solutions provided by citizens could be an option to fill these gaps in mobility supply in the countryside. In order to integrate ride-sharing offers into multimodal travel chains consisting of all available means of transport, the research project NEMo developed a comprehensive mobility platform including the mobile application 'Fahrkreis', functioning as a mobility assistant. Individual preferences such as travel time, monetary costs and environmental impact are considered when searching for the best mobility option. The mobile application integrates diverse mobility offers such as public transport like tram, bus or railway and car-sharing services. As a special innovation provided by NEMo, private ride-sharing options are integrated to fill in supply gaps. The idea is that the community of citizens is becoming an additional mobility supplier which is enriching the offer of local mobility solutions. Car passengers can be picked up at bus stops, supermarkets or railway stations especially at off-peak times (at night, in the very early morning, at weekends etc.). An algorithm matches the driver and other passengers and thus allows pooling diverse interests. One of the main goals is to increase the average utilization of private cars in order to decrease environmental impacts. [9]

To implement new mobility solutions that are widely accepted by the public, a deep understanding of the local circumstances is an important requirement [10]. In fact, there is not one but many types of rural areas - as these differ widely in case of mobility demand and supply. Some regions offer more, some offer less attractive public transport systems. Other regions are geographically more suitable for cycling or do not have any connection to freeways. This is also one of the reasons that there are so many mobility applications available (e.g. the Pendlerportal Regionalverband Braunschweig[7], Mitfahrerbank Varel[8], moobil+ ('Bürgerbus')[9]) which differ from district to district or even from village to village. In any case, the support of the local authorities (as the municipal administration and municipal transport services) is quite important to make the ride-sharing solution popular and to reach a critical mass of potential users. In

[7] https://www.regionalverband-braunschweig.de/pendlerportal/
[8] https://mitfahrerbank-varel.de/
[9] https://www.moobilplus.de/

the case of NEMo, the Wesermarsch district, which is based in the very northern part of Lower Saxony/Germany, advertised the 'Fahrkreis' application actively to its citizens. During the lifetime of the project, 'Fahrkreis' has been actively used by about 30 inhabitants (or commuters) from the Wesermarsch district and about 120 users from all over Germany [11]. To attract the highest possible number of people, the application is ideally adjusted to the specific local circumstances (local branding or integration of voluntary mobility offers as a 'Bürgerbus' service by the citizens for instance).

To understand the concrete mobility demands of the citizens living in the Wesermarsch district, social scientists who worked in the NEMo project captured challenges and barriers for the usage of ride-sharing solutions in rural areas. Results of studies and surveys indicate that although many citizens are interested and perceive advantages for themselves as well as for the environment, the willingness to renounce the private car in the countryside is much lower compared to urban areas. Moreover, the social scientists found out that not only young people during job education have a high potential to use ride-sharing options, but also older people during pension time as these groups do not have many scheduled obligations compared to people working full time [12].

Another item investigated by the team of researchers was the motivation for using ride-sharing solutions. Especially financial, ecological and social aspects are key factors regarding the sharing of private mobility options. Not surprisingly, there are also reservations against using ride-sharing solutions due to limitation of own privacy sphere, uncertainty regarding the reliability of the driver or fellow passengers and open questions in matters of law liability. [12] According to a survey with citizens conducted in the project drivers and fellow passengers are facing uncertainties regarding legal liability and insurance (e.g. cancellation of an already booked trip, accidents, pollution of the car interior) [13]. These research results reveal that the NEMo project did not solely implement a technical solution. To understand and to create appropriate approaches regarding the solution of mobility problems in rural areas, a deeper knowledge of problems from different perspectives is necessary. Besides IT professionals and social scientists experts on privacy issues and business models also participated in an interdisciplinary approach to find new solutions regarding mobility problems on the countryside.

1.2. Functionalities of the 'Fahrkreis' application

As a multimodal travel assistant, 'Fahrkreis' supports the user to find the best travel route taking the preferred means of transport and the number of passengers into consideration (see Figure 1, left side). The reporting system (see Figure 1, middle) of the app delivers information about personal monthly CO_2 emissions for drivers as well as for passengers. Thereby, 'Fahrkreis' can determine exact CO_2 emissions by choosing the respective car from a list. The chart in Figure 1, middle, compares personal emissions (green line) with emissions of the average user (blue line). This could lead to competition between the participants. Thereby, the users are motivated to behave in a more sustainable way. In addition, a realistic cost estimation for the active use of car-sharing is provided.

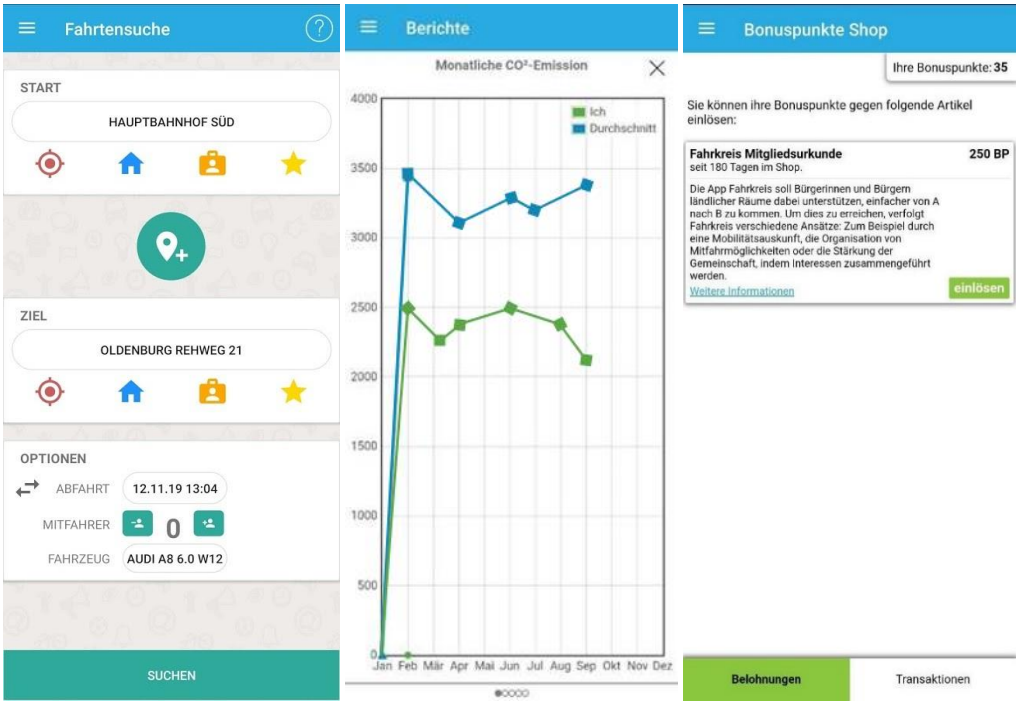


Fig. 1. Landing page (left side), data reporting of individual CO_2 emission levels (middle) and bonus points shop with redeemable items (right side).

To motivate a repetitive usage, 'Fahrkreis' integrates gamification approaches. Moreover, the application awards virtual bonus points for sustainable behaviour. These bonus points can be redeemed for 'real' everyday items in a shop-like environment (see Figure 1, right side). The prerequisite is a cooperation of regional companies or associations and the providers of the mobile application. For example, users could redeem a certain number of bonus points for tickets to the local club or for a cup of coffee or tea at the local coffee shop. In addition, there are many other possible offers that can be integrated into the bonus points shop. [14]

Many of these gamification approaches are already established in real life applications such as the bonus point system of Deutsche Bahn.[10] Loyal clients are awarded by prizes like free tickets, hotel vouchers or the option to donate the bonus points for the good of social projects. Using 'Fahrkreis', there are three ways to collect bonus points. The first option is public transport. The second, and even the most environmentally friendly option, is collecting bonus points by using the bicycle. The third option to collect bonus points is using ride-sharing opportunities via mobile application. Moreover, it is possible to adjust the weighting of the three options and thereby motivate citizens to prioritize the most environmentally friendly option. 'Fahrkreis' is available in common app stores like Google Play Store.[11]

[10] https://www.bahn.de/p/view/bahncard/bahnbonus/bahnbonus_praemien_uebersicht.shtml
[11] https://play.google.com/store/apps/details?id=de.uol.fahrkreis

1.3. Continuation of the research results in real life organizations

The NEMo project ended with a publicly accessible final event on the 13th of February 2020 at the University of Oldenburg. More than 140 participants from citizenry, municipal administration, business and research were informed about the key results of the project [15]. When the duration (and the funding) of research projects like NEMo comes to an end, the question how to continue the results in real life practice becomes important. Therefore, it is necessary to develop concepts for consolidation. A promising strategy could be to implement the technical solutions from the project in real life businesses and organizations to further test and evaluate these. This will then lead to more knowledge about the potential of the mobility platform and 'Fahrkreis'. Aiming at increasing the critical mass of active users via creation of new use cases, the ride-sharing approach of NEMo needs to be applied to organizations which are facing mobility problems in real life. The mobility platform could be tested as part of the mobility management of bigger employers in the countryside. In this specific use case, there is often no suitable supply of public transport available. Many employers located in those areas face the problem that too many employees commute to work by private car without picking up colleagues. This behaviour leads to an insufficient amount of parking space, traffic jams, environmental problems and a decreasing attractivity of employers themselves. In order to make ride-sharing solutions available for the workforce, employers need to raise awareness about the potential and advantages of such possibilities among their employees. Incentive schemes as financial advantages or more spare time could be applied. These could then be based on the bonus point collection system mentioned above. Shift workers probably have a higher potential of using ride-sharing solutions compared to workers with flexible working time as the beginning and end of their working time is more predictable [16].

2. Potentials of ride-sharing solutions in real time

An open research question not considered as part of the NEMo project is how ride-sharing could be realized spontaneously even after the start of a trip. And yet this question is quite relevant when considering the promotion of ride-sharing solutions in the mobility management of businesses as described in the previous section. An employee who missed the morning bus, for instance, could be a target user. Moreover, an employee without a driving license, but a sick husband or wife would need a ride-sharing option on a very short-term basis. Making ride-sharing more flexible and spontaneous is one of the main goals of the recently started research project ***instaride (Development of an innovative ride-hailing solution)***[12].

In general, instaride counts on the potential of everyday rides that happen anyway – e.g. to reach the workplace or to get to university. These rides pass by numerous destinations for which there are potentially fellow passengers. Among available solutions on the ride-sharing market, the higher planning effort only pays off for longer journeys. With the help of an instant-matching algorithm, sharing and travelling together over intermediate and short distances becomes attractive. instaride aims at connecting commuters and fellow passengers efficiently, safely, and instantly by reducing their costs and decreasing the amount of traffic. Latter leads directly to a reduction of CO_2-emissions. Caulfield gauged that through individuals' ride-sharing up to 12,674 tons of such emissions could be saved in Dublin annually [17].

[12] http://www.instaride.de

Furthermore, the instaride application enriches a spontaneous ride-sharing solution with features which are known from social networks. In order to tackle reservations against using ride-sharing solutions as limitation of the privacy sphere or uncertainty regarding the reliability of the driver, instaride contains a whitelist feature. Basically, the whitelist works as a friends list. Within the app, users can search for relatives, acquaintances, friends, or colleagues and send a whitelist request. If the recipient confirms the request, recipient and sender are in the respective whitelists. This aspect facilitates advantages for the driver as well as for fellow passengers.

The following explanations illustrate these and are based on an example in which a driver takes a ride from Bremen to Hamburg. In this case, the driver uses the instaride app and is willing to take fellow passengers on board. During the journey, the driver receives a message that a passenger wants to join in the ride for a certain section. Thereby rating, profile picture and name of the passenger as well as the revenue of the trip are displayed to the driver via the vehicle's infotainment system. The shared journey only takes place if the driver explicitly accepts the requesting passenger. In case of acceptance, the driver gets navigated to the meeting point. At the same time, the passenger gets navigated to the meeting point via his or her instaride app. Within this process, it is planned to open a communication channel as for example in form of a voice call option or chat function.

The whitelist feature plans to speed up this process. If the driver and passenger are in the respective lists, the request of the passenger will be accepted automatically. This means that an explicit authorization by the driver is not required anymore. It is assumed that this feature promotes the use of instaride because it builds trust and reduces reservations. Kim et al. found out that social networking features generally improve perceived value and satisfaction of users and thereby drive on-going mobile user engagement [18]. Furthermore, instaride targets the integration of a related function called blacklist. With the help of the blacklist it would be possible to block people based on bad experiences from former trips. In general, it would also be possible to report users who have behaved improperly during the trip easily via app.

Regarding the mobility management of businesses, instaride opens the possibility to enable ride-sharing on a very short-term basis through a very flexible and spontaneous approach. Taking the social perspective into account, the whitelist feature also supports networking among colleagues and helps to strengthen existing relationships [18].

3. How to integrate local limited mobility platforms in one holistic solution?

Many local initiatives by rural districts, communities and research projects work on their own locally limited shared mobility projects. In the end, this leads to competing solutions with similar technical approaches but a limited scope. The following explanations are based on a fictional business trip from Oldenburg to Berlin. This specific mobility use case is currently not well supported by the ‘Fahrkreis’ application of the NEMo project because it lacks a critical mass beyond the Oldenburg and Wesermarsch regions to cover the total trip. This emphasizes the need for integration solutions like a holistic digital mobility marketplace which combines different locally limited mobility platforms. Such marketplace makes multiple registrations on different platforms obsolete because only a single registration for accessing the marketplace is necessary. Afterwards each integrated solution is usable without any restrictions. Furthermore, the mobility marketplace contributes to reach the critical mass of app users that is necessary

to achieve a minimum level of mobility service reliability. Moreover, the holistic marketplace generates a smooth transition from one local mobility app to another hence the business trip does not end at borders of districts, federal states or even countries.

There are already many well-working mobility solutions available on the local level as the above mentioned 'Fahrkreis' app in the Oldenburg area, 25ways[13] in Hamburg or MobiDig[14] in the Hochfranken region. If these solutions would be integrated in a comprehensive mobility marketplace, the respective applications could be connected and combined to further expand their individual sphere of activity. Despite that, the integration of diverse apps into one holistic solution could face some challenges. On the one hand, there might be different economic interests by the providers involved that complicate the integration process. On the other hand, it may be difficult to connect the different software solutions due to compatibility issues. Therefore, the definition of standardized application programming interfaces (APIs) that enhance the connection of the different shared mobility solutions is a crucial factor for the success and the quick distribution of the holistic mobility marketplace.

4. Outlook: How to extend mobility platforms to strengthen communities in rural areas?

On one side, mobility platforms can make an important contribution to increase the attractiveness of rural areas and on the other side, they enable new social contacts. However, living together in rural areas is not limited to mobility alone, but also to other fields of social life that are essential for the formation of a local community. These include e.g. health, culture, mutual support, social care or common activities. Demographic change, which is one of the key problems on the countryside, means that more people need support and are even more dependent on the community. Therefore, mobility platforms need to be transformed to community platforms which further strengthen cohesion of the citizenry. Furthermore, platforms as Zwopr[15] can help to bolster interpersonal relationships in rural areas. Zwopr is providing an app to connect registered users to exchange help services among each other.

Moreover, integrated solutions can include other non-mobility-related services. Those services could incorporate offers by citizens to support the neighborhood, private rental of tools or technical equipment, gardening, support with garbage disposal, walking the dog, taking the car to the garage, taking care of children, community cooking, coordination of donations and support for the homeless, private collection of clothing and food donations, joint events plus walks and bike rides together. Ride-sharing solutions could also be expanded to run errands for friends as buying food, collecting medication from the pharmacy, or picking up relatives from the train station (see as an example nebenan.de[16]).

Related research shows that citizens are generally very interested in participating in community solutions and citizen science approaches if personal benefits become clear. An example is the city of Zwolle, capital of the Province of Overijssel and voted as the smartest city in 2016 in the Netherlands, where the inhabitants are very engaged to participate in citizen science projects. One target is to work together on solutions facing

[13] https://www.25ways.de/
[14] https://www.mobidig.cloud/
[15] https://zwopr.com/
[16] https://nebenan.de/

climate change. As part of the SensHagen project[17] 50 households, authorities and companies were equipped with partly self-built measurement stations based on sensor systems to monitor temperature and humidity levels. In future this approach should be applied to other future topics as livability of public spaces or energy transition. [19] Another example is the Oldenburg based research project ECOSense[18]. Within the project bicycle sensors were provided to citizens in and around Oldenburg. The gathered data base should be used to learn more about demands of cyclists and conditions of road surfaces to improve overall bicycle infrastructure – a target which is easy to understand and many cyclists do want to support actively. After a few days, several hundred citizens from the Oldenburg area registered and wanted to be part of the project. The personal benefit for the citizens was very clear, which led directly to a very high willingness of the people to participate.

Generated results regarding citizenry mobility of the NEMo project can additionally be integrated in new research projects that focus on citizen science approaches as the project ***PUUK (Portal for the demand-oriented provision of environmental information by companies and municipalities with public participation)***[19]. PUUK aims at providing environmental information according to specific needs of citizens and if necessary, visualises their personal mobility information on the portal. Thereby, participants can learn more about their individual behaviour regarding sustainability.

5. Conclusion

NEMo and other projects in the field of sustainable mobility proved that ride-sharing solutions with focus on rural areas have great potential. Several quite well-working solutions have emerged from various projects. Existing solutions could be updated with more functions to make ride-sharing easier and more social as the whitelist function of instaride which supports reaching a critical mass of users. However, most of the mobility platforms stay limited to a small spatial application space. Therefore, these newly developed mobility platforms need to be interlinked as part of integrated solutions to increase the personal range. In order to enable mobility solutions to have a broader field of application and thus a greater opportunity for citizens to participate, other areas of public life that further strengthen the community should gradually be integrated.

Acknowledgments

This research paper was submitted as part of the research projects ***"NEMo - Sustainable fulfillment of mobility demands in rural areas", "instaride - Development of an innovative ride-hailing solution"*** and ***"PUUK - Portal for the demand-oriented provision of environmental information by companies and municipalities with public participation"***.

The NEMo project was funded by the Ministry of Science and Culture of Lower Saxony and the Volkswagen Foundation as part of the funding program "Lower Saxony Vorab" (funding number VWZN3122). The project with a total duration of four years ended at the end of March 2020. A continuation

[17] https://senshagen-zwolle.opendata.arcgis.com/
[18] https://ecosense.mein-dienstrad.de/
[19] https://puuk-projekt.de/

of the research activities and a transfer to practice are already being planned and partly realized (e.g. instaride). A special expression of gratitude goes to the pilot region, the Wesermarsch county, and the large number of associated partners involved in NEMo. Stakeholders from the business world and various associations made an important contribution to the transfer and evaluation of the ideas and applications developed in the NEMo context.

The instaride project is based on the results of the above mentioned NEMo project and is funded by the European Regional Development Fund through the project sponsor NBank (Investment and Promotional Bank Lower Saxony) as part of the "Innovation funding program for research and development in companies" (funding number ZW3-85031028).

The PUUK project plans to create an intermediary portal for visualising environmental information to different stakeholders (citizens, companies and government) and is funded by the German Federal Environmental Foundation Deutsche Bundesstiftung Umwelt DBU (project reference number 35152/01).

References

[1] Küpper, P. (2016): Abgrenzung und Typisierung ländlicher Räume. Braunschweig: Johann Heinrich von Thünen-Institut.

[2] Statistisches Bundesamt (2019): Ausstattung privater Haushalte mit Fahrzeugen - Deutschland. Available online: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Ausstattung-Gebrauchsgueter/Tabellen/liste-fahrzeuge-d.html#fussnote-1-115512. In: destatis. Uploaded 2019. Accessed 7th of July 2020.

[3] Forbes, P. J., Wells, S., Masthoff, J., Nguyen, H. (2012). Superhub: Integrating behaviour change theories into a sustainable urban-mobility platform. In: The 26th BCS Conference on Human Computer Interaction 26, 1-4.

[4] Masuch, N., Lützenberger, M., Keiser, J. (2013). An Open Extensible Platform for Intermodal Mobility Assistance. ANT/SEIT, 19, 396-403.

[5] Lamberti, R., Fries, C., Lücking, M., Manke, R., Kannengießer, N., Sturm, B., ... & Sunyaev, A. (2019). An open multimodal mobility platform based on Distributed Ledger Technology. In: Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer, Cham, 41-52.

[6] Hevner, A. R., March, S. T., Park, J., Ram, S. (2004). Design science in information systems research. MIS quarterly, 75-105.

[7] Peffers, K., Tuunanen, T., Rothenberger, M. A., Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.

[8] Schlump, C. (2018): Verkehrswende auf dem Dorf? Mobilitätsoptionen in ländlichen Räumen. In: Ökologisches Wirtschaften, 2018(2), 23.

[9] Marx Gómez, J. (2020): Nachhaltige Erfüllung von Mobilitätsbedürfnissen im ländlichen Raum - Abschlussveranstaltung. Oldenburg. Available online: https://nemo-mobilitaet.de/blog/wp-content/uploads/2020/02/02_NEMo_Ergebnisse.pdf. In: NEMo. Uploaded 20st of February 2020. Accessed 7th of July 2020.

[10] Becker, J., Hofmann, D., Blees, V., Walther, S. (2017). Abschlussbericht des Forschungsprojekts „MoLa.opt – Mobilität auf dem Land optimieren“. Available online: https://www.frankfurt-university.de/fileadmin/standard/Hochschule/Fachbereich_1/FFin/Neue_Mobilitaet/Veroeffentlichungen/2017/A

bschlussbericht_MoLa.opt_Textteil.pdf. In: Frankfurt University of Applied Sciences. Uploaded 2017. Accessed 12th of July 2020.

[11] Marx Gómez, J. (Ed.) (2020): NEMo – Nachhaltige Erfüllung von Mobilitätsbedürfnissen im ländlichen Raum – Abschlussbericht. Oldenburg, Vechta, Braunschweig, Passau. Available online: https://nemo-mobilitaet.de/blog/wp-content/uploads/2020/08/20200823_NEMo-Abschlussbericht.pdf. Uploaded 3rd of September 2020. In: NEMo. Accessed 4th of September 2020.

[12] Marx Gómez, J. (Ed.) (2018): NEMo – Nachhaltige Erfüllung von Mobilitätsbedürfnissen im ländlichen Raum – Zwischenbericht. Oldenburg, Lüneburg, Braunschweig. Available online: http://nemo-mobilitaet.de/blog/wp-content/uploads/2018/05/201805_NEMo-Zwischenbericht.pdf. Uploaded 20st of May 2018. In: NEMo. Accessed 7th of July 2020.

[13] Sandau, A. (2019). ZENARiO: Kolloquium zur Nachhaltigen Raumentwicklung. Vortrag am 25.10.2020 an der Universität Oldenburg.

[14] Sandau, A., Dietrich, B., Akyol, A., Wagner vom Berg, B. and Marx Gómez, J. (2018). Steigerung der Sensibilität für nachhaltige Mobilität durch die mobile Reiseapplikation Guyde, Tagungsband Multikonferenz Wirtschaftsinformatik 2018). In: Drews, P.; Funk, B.; Niemeyer, P. und Xie, L. (Ed.): Tagungsband Multikonferenz Wirtschaftsinformatik 2018, Lüneburg.

[15] Sandau, A. (2020). NEMo-Abschlussveranstaltung (13.02.2020). Available online: https://nemo-mobilitaet.de/blog/de/nemo-abschlussveranstaltung-13-02-2020/. In: NEMo. Uploaded 14th of February 2020. Accessed 20st of July 2020.

[16] Gerike, R. (2019). Einfluss der Arbeitszeiten auf die Fahrgemeinschaften im Berufsverkehr. Available online: https://www.forschungsinformationssystem.de/servlet/is/112553/. In: Forschungsinformationssystem. Updated 21st of June 2019. Accessed 19th of July 2020.

[17] Caulfield, B. (2009). Estimating the environmental benefits of ride-sharing: A case study of Dublin. Transportation Research Part D: Transport and Environment, 14(7), 527–531.

[18] Kim, Y. H., Kim, D. J., & Wachter, K. (2013). A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention. Decision Support Systems, 56, 361–370.

[19] Dhingra, A. (2019). In Zwolle, people and machines work together for social good. https://www.geospatialworld.net/blogs/in-zwolle-people-and-machines-work-together-for-social-good/. In: Geospatialworld. Uploaded 8th of December 2019. Accessed 19th of July 2020.

Blockchain-based Electronic Record Books for Transparency to Prevent Marine Pollution

Hauke Precht[1], David Saive[2], Simon Czapski[3], Jorge Marx Gómez[4]

1. Introduction

Although, international shipping is subject to many international regulations and numerous port states' measurements towards environmentally friendly shipping, illegal oil and waste dumping is still present. Pollution caused by oil spills is still one of the major marine pollutions from vessels [1]. While the coast guards' activities to combat these unlawful actions are increasing [2], the shipowners' creativity to find ways of illegal dumping increases as well. One phrase commonly used to describe a particularly sophisticated method of illegal oil dumping are so-called "magic pipes" [3]. These are pipes linked to the machinery of a ship to bypass the ship's pollution control systems to illegally discharge oil and oily waters directly to the sea [3]. This is not an uncommon way of circumventing international regulation. For example, in June 2019 a Portuguese shipping company plead guilty to use such a magic pipe [4], and in July 2017 a Greek shipping company was found guilty of the same offence [5]. Such illegal dumping of waste is a significant threat to the environment. One of the biggest cruise shipping companies in the world has been fined 20 million Dollar due to illegal dumping of plastic in The Bahamas [6]. All cases had one thing in common: falsification of the record-keeping, i.e. of oil and other waste record books. As the person responsible for the record keeping, either the captain or the vessel's machinery officer [7], enters every relevant event by hand, they decide if and how a certain pollution-related event is entered into the respective record books. This enables easy counterfeit and fraud considering pollution related record books. Therefore, international regulation towards greener shipping cannot succeed as they lack clear evidence. If falsification of such record books could be prevented, regulations could be applied more strictly.

Nevertheless, international regulation strongly relies on the record-keeping of the ship's emissions. All ships are subject to the "International Convention for the Prevention of Marine Pollution from Ships" (MARPOL) [8]. MARPOL itself is subdivided into a main part and six annexes in which emissions and pollutions are described: oil (Annex I) [9], noxious liquid substances (Annex II) [10], other harmful substances (Annex III) [11], sewage (Annex IV) [12], garbage (Annex V) [13] and Sulphur (Annex VI) [14]. All annexes impose the duty of record-keeping on the shipowners, e.g. regulation 17 and 36 in Annex I for the Oil Record Book Part I [15] and Part II [16]. Part I covers machinery operations while Part II is for cargo/ballast operations. This creates a complete record of all activities concerning oil and oily waters.

MARPOL offers strict guidelines for the form of such record books, e.g. in Appendix III to Annex 1 [17]. To the current day, paper-based record books are still used, although the MARPOL itself allows the

[1] Carl von Ossietzky University of Oldenburg, Oldenburg Germany, hauke.precht@uol.de
[2] Carl von Ossietzky University of Oldenburg, Oldenburg Germany, david.saive@uol.de
[3] Carl von Ossietzky University of Oldenburg, Oldenburg Germany, simon.czapski@uol.de
[4] Carl von Ossietzky University of Oldenburg, Oldenburg Germany, jorge.marx.gomez@uol.de

usage of electronic record books: In 2019 numerous resolutions have been adopted to implement the "Guidelines for the use of electronic record books under MARPOL" [18]. They aim to "facilitate port state control inspections" and "to reduce administrative burdens and contribute to onboard environmental initiatives, e.g. reduction of paper use." [18], since the maintaining of the traditional handwritten oil record book, for example, can be "a real headache" [19], next to the already described issues concerning counterfeit of data. The flag states approve the usage of certain technology for electronic record books [20]. All record books serve to control the correct handling with the named oil, waste or garbage. The entries contained are used as evidence for public inspections as well as in court proceedings [21].

In Germany, electronic record books are governed by the "Bundesamt für Schifffahrt und Hydrographie" (engl. "Federal Maritime and Hydrography"; abbr. BSH). The BSH approves such technology for electronic record books, that does comply with ISO 21745. The before described regulations and guidelines allow the usage of electronic record books in a technology-neutral way. This paper aims to answer the question, if a digitization of a ship's record book benefits from using blockchain as a tamper proof technology. As different countries may use different standards to approve such technology, the focus of this paper is for Germany, i.e. the solution must comply with ISO 21745. The following paper is structured as follows: First, an overview of related work is given. Next, it is evaluated if blockchain is suitable for digitizing record books followed by a proposed solution combining blockchain and IoT sensors. The paper concludes by summarizing the key finding and provides an outlook of next steps and future research.

2. Related Work

Koss already described in 1996, the importance of optimizing waste management within the maritime sector [22]. A general investigation of using distributed ledger technology (DLT) within the shipping industry was carried out in [23], identifying that the usage of DLT for digitizing paper contracts with potential for the scientific use of data, improve shipping operations and processes while making operations safer and environmentally friendly [23]. Similar research is shown in [24] describing existing projects in the maritime sector leveraging blockchain technology, for example, a new verified gross mass of packed containers. Further, the authors describe the possibility of using blockchain for bill of ladings [24] which is also proposed by Wunderlich and Saive [25]. In general, the maritime industry requires external motivation to adopt new technology, for example the possibility to save costs or new legal requirements [24].

Since the ship's logbooks serve to keep track of all types of oil and waste disposal, general approaches to the use of blockchain in waste management must also be considered. Several papers have been identified showing scenarios in which the usage of blockchain technology can be beneficial for waste management. Ongena et al. provides a analysis of the general applicability of blockchain in waste. They identified key problem areas in waste management such as fraud and manipulation, wrong or loss of information, manual processes, lack of knowledge about technology and lack of control [26]. A more specific investigation of blockchain in waste management was carried out by Laouar et al. as they propose to use blockchain for tracking movements and collection of solid waste [27]. The goal of their smart-contract-based system is to enhance the transparency of waste management information [27]. Like the problem areas in waste management in general, identified by [26], Laouar et al. identified seven problems within the current waste

chain and stating cheating, manipulation, loss or wrong information as well as lack of control which were also identified in [27] as key problem areas. Gupta and Punam also propose a smart-contract-based waste management system, focusing on e-waste and regulation constraints in India [28]. A similar approach is presented in [29], where a deposit system, based on blockchain is presented, also focusing on Waste Electrical & Electronic Equipment (WEE). Another waste management system is proposed in [30] motivated by smart cities. Further, they use Unified Modeling Language (UML) and Temporal Logic of Actions (TLA+) to proof correctness preciseness and completeness of the proposed modelling [30]. Hakak et al. add another point focusing specifically on wastewater management. They propose to tackle current shortcomings of tampering or modifying sensor data, currently stored in databases, by leveraging blockchain technology [31]. An approach for cross-border waste tracking is shown in [32]. The authors describe a system based on blockchain for tracking cross-border movements and preventing illegal dumping while maintaining data protection standards [32].

As shown, there is progress in the use of block chain technology in the maritime sector and in waste management. However, to best of the authors knowledge, no scientific work is dealing with blockchain based digitization of a ships record book. Therefore, this paper aims to answer the general question, if the digitization of a ships record book based on blockchain is beneficial and if this meets requirements implied by ISO 21745 in order to improve data quality and to counter manipulation of record books to disguise illegal waste dumping on sea.

3. Towards a Blockchain and IoT based Record Book

Even though no scientific literature could have been identified, several solutions for digital record books exist within the maritime industry [33]. But only one solution, RINA Cube, advertised the possibility to use a blockchain-based oil record book [34]. However, no information on the used technology or actual usage could be acquired, as the authors did not receive answer upon contacting the company. That way, questions concerning privacy, reliability of data, regulatory compliance and general architecture remain unclear. Therefore, it is not possible to determine which flag state would approve this solution. Also, the presented solutions rely on manual entering of data, still opening the gate for counterfeit. This leads to the need of a general and transparent approach for digitizing the ship record book based on blockchain while solving problems resulting from entering data manually. Also, as mentioned, the general question if blockchain is generally suitable for digitizing a ship's record book needs to be answered as well. Further, it must be ensured that such solution is legally complaint, leading to the need to integrate law researchers early in the process, resulting in the need for interdisciplinarity.

As a first step, it is necessary to examine whether the usage of blockchain is beneficial when digitizing a ship's record book. To do so, it is first discussed, if the identified problem areas from general waste management (based on [26]), which can be optimized by blockchain, also apply to the maritime sector, specifically the ship's record book. If these problem areas can be transferred to this application case, it can be argued that a block chain is also advantageous, as it has already proven to be beneficial in view of the main problem areas of general waste management [26]. Therefore, each in [26] identified problem area, *fraud manipulation*, *wrong or loss of information*, *manual processes*, *lack of knowledge about technology* and *lack of control* is discussed in terms of ship record books.

Considering the first identified problem area, fraud manipulation, the authors state that entries can easily be counterfeit or afterwards modified [26]. As the ship's record book is paper-based, similar problems are present. It is stated in [26], that blockchain technology is not suitable to counter this issue, as the initial entry could be already wrong, even though the blockchain permits modification. To tackle this issue, Internet of Things (IoT) sensors could be applied. IoT describes the concept to allow several objects to collect and exchange data via network [35]. Such IoT sensors could be used to eliminate a possible source of error, i.e. the manual entering of data, by producing the corresponding data and storing it in the blockchain. A similar concept of combining IoT sensors and blockchain is already well known in supply chain traceability, as shown for example in [36, 37, 38]. In terms of general waste management tracking, Yeong et al. proposed an NFC-based waste management tracking system [39] whose concepts might apply to data gathering for ship record books. Closer related to electronic record books are the current attempts of connecting sensors in the bunker holds for automatic bunker (quality) tracing [40]. However, the usage of IoT sensors does not eradicate possible fraud manipulation completely. This is due to the nature of IoT architecture in which the sensors usually send data to (central) gateways for further processing, e.g. sending it to a blockchain. These (central) gateways could be attacked or be corrupted by the operator. Latter can be tackled in a way, that such gateways are not operated by ship owners but by a governmental or independent institution, leading to a counter-measurement on an organizational level. As the problem of data integrity is a problem for IoT or cyber-physical systems (CPS) in general, dedicate research areas emerged. A promising approach is to add a digital watermark to the generated data of the sensor [51]. The sensor itself would add this watermark, so manipulating data send from those sensors will be detectable, leading to a reliable and hard to manipulate data source. Those approaches must be further evaluated and integrated into the proposed solution towards a secure system in general, which is out of scope for this paper.

Analyzing the second identified problem area, wrong or loss of information, it shows that this issue also occurs at ship record books, as entered data could be counterfeit or the whole book could be lost. That would be, in the worst case, if the ship sinks. As blockchain provides a decentral and distributed environment which stores the data, this technology is suitable to counter this identified problem [26].

The third problem area describes the manual process, as data needs to be manually gathered and entered in paper-based records, which is also done in ship record books. This cannot be solved by blockchain technology alone [26], but as stated above, the combination of IoT concepts and blockchain could automate the whole process of keeping records for ship record books which would save time for the crew.

Based on the conducted interviews in [26], a lack of knowledge and ability to work with technology within the waste management sector is shown. Similar can be observed in the maritime sector, which is slow to adopt new technology and requires external motivation, e.g. by new legal requirements or potential cost/time saving [26]. This is a problem, which cannot be solved by blockchain or technology in general but must be done step by step, supported by change management methods.

The last identified problem area by Ongena et al. describes the lack of control, describing the issue that governmental inspection at a waste division stations are time-consuming [26]. This can be simplified and solved by a blockchain-based solution serving as a trust factor, given that the organization can ensure that the data is entered correctly [26]. This also applies to inspections of ship record books. The use of IoT sensors can meet the requirement that the correct input of data must be guaranteed.

As shown, the general problem areas in waste management process also apply to ship record books. Thus, blockchain-technology should be applied. By combining IoT, i.e. sensor, and blockchain technology, problems considering data integrity can be solved. As the paper aims to show that a digital record book based on blockchain, now enhanced with IoT sensors, is usable in Germany, it is evaluated next if such solution is compliant with ISO 21745 as this is a condition to be met for acceptance by the BSH. This analysis is shown in Table 1. where the five fundamental functional requirements based on ISO 21745 are shortly described, and the proposed solution is discussed in terms of fulfilment.

ISO 21745 Requirement	**Blockchain and IoT Solution**
Data Storage: The data stored in an electronic record book must be identical with the paper-based record books. UTC as time format must be used, and all information must be in a clear legible font. Four types of data are supported: automatically-collected data, record book data, signed record book data and edit history data. [41] Further, it is described in that input values need to be protected, preventing manipulations and counterfeit along with an authorization system [41]. It is also required to provide monitoring of the system along with the possibility of status reports [41].	Using blockchain technology and sensor-generated data does not limit any of these requirements. Further, automatically-collected data is explicitly supported, thus enabling the usage of aforementioned IoT concepts. The requirement of an authorization-based system where users can be identified can be fulfilled by applications on top of the actual blockchain or by using private permissioned blockchains [42] with specific governance.
Record management: In the norm, it is stated, that automatically-collected data shall not be revisable and modifiable while record book data can only be modified by authorized persons. Further, changing actions need to be tracked and made visible. Note, that it is explicitly written, that records need to be protected against unauthorized deletion, destruction or amendment and the system has to take care of measurements against unauthorized or untraceable changes [41]. Further, the system shall store edit history data, like contents time, place and user who made the change. For automatically generated data, it is written that it shall not be revised and modifiable [41]. In terms of data synchronization, similar requirements apply, i.e. the content must remain unchanged.	Leveraging blockchain technology, entered data can be considered tamper-proof and non-modifiable [43, p. 125], fulfilling the stated requirement. Further, transaction data saves the required information of who has called the function, and what record entry has been made. Due to the decentralized and distributed nature of blockchain, it not possible to destroy or destruct the system, which fulfils another requirement in terms of record management. Similar holds true considering automatically generated data, which is part of the proposed combination of IoT sensors and blockchain for digitizing the ship record book. Also, the last stated requirements that the data remains unchanged when synchronizing is inherited by the nature of blockchain.
System output: This requirement describes that the output contents of the electronic record book	When using blockchain technology, the data is stored in blocks. An application layer is required to

must meet the requirements by the flag state [41], e.g. in this example, the requirements of Germany. Further, a file format shall be used, which prevents modification or editing, e.g. PDF [41].	provide functionality to the end users [43, p. 19], e.g. in this case a PDF export. This is similar to a classic software stack using relational databases. It can be stated that this requirement can be fulfilled when using blockchain, as it does not affect the storage of data in any way but only the presentation of the data.
Validation: The electronic record book shall provide and audit logging, stating what has been done and by whom and at what time [41]. Therefore, a user and authentication system is required along with role-based access control while also providing approval route [41]. Further, it is explicitly stated that the contents can be entered either automatically or manually.	The mentioned requirements in terms of validation can be linked to already described requirements in record management. Enhancements to the requirements concerning audit logs is the role-based access control. This is an aspect which needs to be considered when selecting the actual blockchain implementation. A private permissioned blockchain might provide a more sophisticated permission system than a public one, where such a system might be built around the actual blockchain. Nevertheless, it is no knockout criterion which prevents the usage of blockchain technology. Also, it is again stated, that records can be entered automatically, clearing once again the path for the usage of IoT.
System availability: This requirement states that it shall be possible to access and to create new records even if a storage medium fails, so at least two independent storage mediums are required [41].	As stated, the decentral and distributed nature of blockchain enables a high fault tolerance while keeping the system available. Therefore, the requirement of system availability can be achieved when using a blockchain approach.

Table 1. ISO 21745 compliance for blockchain combined with IoT.

Note that due to space limitations of the paper, only the key functional requirements of ISO 21745 were analyzed. The norm further describes factors like human-machine interface, the handling of system updates and test methods to verify the mentioned key functional requirements [41]. The comparison of requirements based on ISO 21745 and the proposed system consisting of blockchain and IoT sensors shows that it is in general applicable to use such a system while being compliant with the ISO norm. As next steps, a proof of concept-based prototype needs to be developed. It must be evaluated, what kind of IoT sensors could be used to track data automatically and how they could be integrated into the ship's infrastructure. Further, a blockchain (e.g. Ethereum, Hyperledger Fabric,…) needs to be selected, suitable for this specific use case. Questions concerning the scope of access (private vs public) and the used consensus algorithm need to be answered thoroughly. A possible approach for the selection process is described by Precht et al. [44]. In particular, the consensus algorithm should be carefully evaluated from a sustainability perspective, with respect to the high energy demand of proof-of-work algorithms. [45].

4. Conclusion and Outlook

As shown no research considering the applicability of blockchain for record books could have been identified, although progress has been made in the field of waste management in general. However, industry solutions exist. But only one of these solutions is blockchain-based. As no answer on the authors' inquiries were received, it remains unclear what technology is used and if this would be compliant to the ISO 21745 norm, which is required for a usage in Germany. Further, the general question if the process around the record books benefits form the usage of blockchain technology is not yet considered. In this context the paper presents the applicability of the identified problem areas from general waste management to the record book, namely *fraud manipulation*, *wrong or loss of information*, *manual processes*, *lack of knowledge about technology* and *lack of control*. It was also found that with regard to fraud manipulation, the use of blockchain technology alone is not sufficient as it does not prevent the entry of falsified data. To tackle this problem, a combination of blockchain and IoT sensors is proposed which automatically enters the respective data. As this paper is intended to show a possible applicability in Germany, the combination of blockchain and IoT sensors is evaluated for compliance with the requirements of ISO 21745 and it is shown that the proposed solution meets these requirements.

The distributed architecture of the blockchain and its resilience towards changes of the stored data requires a strict analysis of the usage of personal data [46]. Therefore, compliance with the requirements of ISO 21745 alone is not sufficient for the legally compliant use of block-chain based verification books. The General Data Protection Regulation (GDPR) [47] imposes various duties on the processor, i.e. offering ways for the rectification [48] or deletion of personal data [49]. If more than one company has access to the ledger, antitrust-regulations must be considered [50]. Besides, technical questions, such as which sensors should be used, how they should be installed on the ship and the general architecture, must be answered. This could be done by using a proof-of-concept approach. If all these challenges are met, blockchain-based electronic record books using IoT sensors can help improve data quality and thus help to enforce international regulations for more environmentally friendly shipping and transportation.

References

[1] Karim, Md S. (2015): Prevention of Pollution of the Marine Environment from Vessels. The Potential and Limits of the International Maritime Organisation, 44.

[2] Gullo, Benedict S. (2011): Illegal Discharge of Oil on the High Seas: The U.S. Coast Guard's Ongoing Battle against Vessel Polluters and a New Approach Toward Establishing Environmental Compliance. Military Law Review, vol. 209, no. 1, Fall 2011, 122-185.

[3] Pasfield, B.; Rindfleisch, E. (2009): Finding the magic pipe: Do seamen have constitutional rights when U.S. coast guard boarding turns criminal. University of San Francisco Maritime Law Journal, 22(1), 23-38.

[4] The United States Department of Justice (2019): https://www.justice.gov/opa/pr/portuguese-shipping-company-pleads-guilty-falsifying-oil-record-book-and-obstruction (07.07.2020).

[5] MFAME (2017): https://mfame.guru/guilty-falsifying-ships-oil-record-book/ (07.07.2020).

[6] Forbes (2019): https://www.forbes.com/sites/trevornace/2019/06/11/carnival-cruise-to-pay-20-million-after-admitting-to-dumping-plastic-waste-in-the-bahamas/#417e77153a6f (07.07.2020).

[7] Mura, Jennifer (2018): Oil Pollution Violations and Enforcement: Who is responsible for Maintaining the Oil Record Books, 17 Lloyds Maritime Law Journal 2018, 381-408.

[8] International Convention for the Prevention of Pollution from Ships, opened for signature 2 November 1973, 12 ILM 1319 as modified by the Protocol of 1978 to the 1973 Convention, opened for signature 17 February 1978, 1341 UNTS 3 (entered into force 2 October 1983) (MARPOL 73/78).

[9] MARPOL Annex I Regulations for the Prevention of Pollution by Oil (entered into force 2 October 1983).

[10] MARPOL Annex II Regulations for the Control of Pollution by Noxious Liquid Substances in Bulk (entered into force 2 October 1983).

[11] MARPOL Annex III Prevention of Pollution by Harmful Substances Carried by Sea in Packaged Form (entered into force 1 July 1992).

[12] MARPOL Annex IV Prevention of Pollution by Sewage from Ships (entered into force 27 September 2003).

[13] MARPOL Annex V Prevention of Pollution by Garbage from Ships (entered into force 31 December 1988).

[14] MARPOL Annex VI Prevention of Air Pollution from Ships (entered into force 19 May 2005).

[15] MARPOL Annex 1, Regulation 17 - Oil Record Book, Part I - Machinery space operations.

[16] MARPOL Annex 1, Regulation 36 - Oil Record Book, Part II - Cargo/ballast operations.

[17] MARPOL Annex 1, Appendix III: Form of Oil Record Book.

[18] IMO Marine Environment Protection Comittee (2019): Resolution MEPC.312. (74) adopted on 17th of May, 2019.

[19] Stefopoulou, E. (2018): Avoiding an onboard record-keeping headache key for smooth sailing. In Liz Booth. Maritime Risk International, 20th July 2018. London. Informa, 2018.

[20] Resolution MEPC.312 (74), Sec. 1.4.

[21] De La Rue, C. (2009): Shipping and the Environment. 2nd Edition. London. Informa, Chapter IV.

[22] Koss, L. Technology development for environmentally sound ships of the 21st century: an international perspective. J Mar Sci Technol 1, 127–137 (1996). https://doi.org/10.1007/BF0239117.

[23] Vujičić, Srdjan; Hasanspahić, Nermin; Car, Maro; Čampara, Leo (2020): Distributed Ledger Technology as a Tool for Environmental Sustainability in the Shipping Industry. In: JMSE 8 (5), S. 366. DOI: 10.3390/jmse8050366.

[24] Czachorowski K., Solesvik M., Kondratenko Y. (2019) The Application of Blockchain Technology in the Maritime Industry. In: Kharchenko V., Kondratenko Y., Kacprzyk J. (eds) Green IT Engineering: Social, Business and Industrial Applications. Studies in Systems, Decision and Control, vol 171. Springer, Cham.

[25] Wunderlich, Stefan; Saive, David (2020): The Electronic Bill of Lading. In: Javier Prieto, Ashok Kumar Das, Stefano Ferretti, António Pinto und Juan Manuel Corchado (Hg.): Blockchain and Applications, Bd. 1010. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), S. 93–100.

[26] Ongena, Guido; Smit, Koen; Boksebeld, Jarno; Adams, Gerben; Roelofs, Yorin; and Ravesteyn, Pascal, "Blockchain-based Smart Contracts in Waste Management: A Silver Bullet?" (2018). BLED 2018 Proceedings. 19. https://aisel.aisnet.org/bled2018/19.

[27] Laouar, Mohamed Ridda; Hamad, Zaineb Touati and Eom, Sean. 2019. Towards blockchain-based urban planning: Application for Waste Collection Management. In Proceedings of the 9th International Conference on Information Systems and Technologies (icist 2019). Association for Computing Machinery, New York, NY, USA, Article 39, 1–6. DOI:https://doi.org/10.1145/3361570.3361619.

[28] N. Gupta and P. Bedi, "E-waste Management Using Blockchain-based Smart Contracts," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 915-921, doi: 10.1109/ICACCI.2018.8554912.

[29] Sebastian Lawrenz, Vera Stein, Lukas Jacobs, and Andreas Rausch. 2020. A Blockchain-Based Deposit System to Reduce WEE. In Proceedings of the 2020 The 2nd International Conference on Blockchain Technology (ICBCT'20). Association for Computing Machinery, New York, NY, USA, 130–134. DOI:https://doi.org/10.1145/3390566.3391686.

[30] S. Latif, A. Rehman and N. A. Zafar, "Blockchain and IoT Based Formal Model of Smart Waste Management System Using TLA+," 2019 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2019, pp. 304-3045, doi: 10.1109/FIT47737.2019.00064.

[31] S. Hakak, W. Z. Khan, G. A. Gilkar, N. Haider, M. Imran and M. S. Alkatheiri, "Industrial Wastewater Management using Blockchain Technology: Architecture, Requirements, and Future Directions," in IEEE Internet of Things Magazine, vol. 3, no. 2, pp. 38-43, June 2020, doi: 10.1109/IOTM.0001.1900092.

[32] D. Schmelz, K. Pinter, S. Strobl, L. Zhu, P. Niemeier and T. Grechenig, "Technical Mechanics of a Trans-Border Waste Flow Tracking Solution Based on Blockchain Technology," 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW), Macao, Macao, 2019, pp. 31-36, doi: 10.1109/ICDEW.2019.00-38.

[33] International Registries (2020): Approved electronic record book providers. International Registries. Online at https://www.register-iri.com/maritime/maritime-technical-support/marpol-electronic-record-books/, accessed on 14.07.2020.

[34] RINA (2019): First time ever Blockchian applied to Oil Record Book. RINA. Online at https://www.rina.org/en/media/press/2019/06/17/blockchain, zuletzt aktualisiert am 17.06.2019, accessed on 14.07.2020.

[35] Cui, P., Guin, U., Skjellum, A. et al. Blockchain in IoT: Current Trends, Challenges, and Future Roadmap. J Hardw Syst Secur 3, 338–364 (2019). https://doi.org/10.1007/s41635-019-00079-

[36] M. P. Caro, M. S. Ali, M. Vecchio and R. Giaffreda, "Blockchain-based traceability in Agri-Food supply chain management: A practical implementation," 2018 IoT Vertical and Topical Summit on Agriculture - Tuscany (IOT Tuscany), Tuscany, 2018, pp. 1-4, doi: 10.1109/IOT-TUSCANY.2018.8373021.

[37] S. Malik, V. Dedeoglu, S. S. Kanhere and R. Jurdak, "TrustChain: Trust Management in Blockchain and IoT Supported Supply Chains," 2019 IEEE International Conference on Blockchain (Blockchain), Atlanta, GA, USA, 2019, pp. 184-193, doi: 10.1109/Blockchain.2019.00032.

[38] S. Madumidha, P. S. Ranjani, S. S. Varsinee and P. S. Sundari, "Transparency and Traceability: In Food Supply Chain System using Blockchain Technology with Internet of Things," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 983-987, doi: 10.1109/ICOEI.2019.8862726.

[39] Beh Choon Yeong, Nurul Hashimah Ahamed, Hassain Malim, and Manmeet Mahinderjit Singh. 2017. NFC-based waste management tracking and monitoring system. In Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing (ICC '17). Association for Computing Machinery, New York, NY, USA, Article 86, 1–5. DOI:https://doi.org/10.1145/3018896.3025134.

[40] T. Pitana, E. Kobayashi and N. Wakabayashi, "Estimation of exhaust emissions of marine traffic using Automatic Identification System data (case study: Madura Strait area, Indonesia)," OCEANS'10 IEEE SYDNEY, Sydney, NSW, 2010, pp. 1-6, doi: 10.1109/OCEANSSYD.2010.5603866

[41] Wagner, N., Wiśnicki, B. (2019): Application of blockchain technology in maritime logistics, DIEM Dubrovnik International Economic Meeting, Vol. 4. No. 1, 2019, p. 162.

[42] INTERNATIONAL STANDARD ISO 21745, 21745, 2019.

[43] Mattila, Juri (2016): The Blockchain Phenomenon. The Disruptive Potential of Distributed Consensus Architectures. In: ETLA Working Papers (38). Online verfügbar unter http://pub.etla.fi/ETLA-Working-Papers-38.pdf, zuletzt geprüft am 22.02.2019.

[44] Singhal, Bikramaditya; Dhameja, Gautam; Panda, Priyansu Sekhar (2018): Beginning Blockchain. Berkeley, CA: Apress.

[45] Precht, Hauke; Wunderlich, Stefan; Marx Gómez, Jorge (2020): Applying Software Quality Criteria to Blockchain Applications: A Criteria Catalog. In: Tung Bui (Hg.): Proceedings of the 53rd Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences: Hawaii International Conference on System Sciences (Proceedings of the Annual Hawaii International Conference on System Sciences).

[46] Vries, Alex de (2018): Bitcoin's Growing Energy Problem. In: Joule 2 (5), S. 801–805. DOI: 10.1016/j.joule.2018.04.016.

[47] Janicki, T., Saive, D. (2019): Privacy by Design in Blockchain-Netzwerken, Zeitschrift für Datenschutz, 2019, p. 251-256.

[48] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[49] Art. 16 GDPR.

[50] Art. 17 GDPR.

[51] Louven, S., Saive, D. (2019): Antitrust by design: The prohibition of anti-competitive coordination and the consensus mechanism of the blockchain, GRUR Int. 2019, p. 537-543.

[52] H. Juma, I. Kamel, and L. Kaya, “On protecting the integrity of sensor data,” in 2008 15th IEEE International Conference on Electronics, Circuits and Systems, St. Julien's, Malta, Aug. 2008 - Sep. 2008, pp. 902–905.

PART II
ENVIRONMENTAL INFORMATION SYSTEMS

Towards Decision Tree Based Assistance Functions of a Cloud Platform for Environmental Compliance Management

Heiko Thimm [1]

1. Introduction

Managing compliance with environmental regulations belongs to the core tasks of the corporate work area of environmental management [1]. This work area imposes challenging responsibilities. Among other reasons the challenges arise from the fact that today, the business world faces a high density of environmental regulations that do not remain static. As regulations are added, revised, and withdrawn, companies must maintain accurate compliance records that prove 1. that they are aware of all relevant regulations and 2. that they comply to these regulations through proper enforcement measures.

In our research program we are investigating how the work area of corporate environmental compliance management (CECM) can obtain effective support by a dedicated full-fletched information system that offers rich functions for typical operational and managerial CECM tasks. A first research prototype with a traditional architecture that was limited to an on-premise deployment mode is now being further developed to a CECM cloud platform [2, 3]. It is assumed that a cloud computing paradigm will offer several benefits to small and medium sized (SME) companies. For example, SME companies often need to engage external actors in CECM tasks (e.g. environmental consulting companies, freelancers, agencies, certified operational officers) due to limitations of their resource capacities and limited expertise. Obviously, a corresponding CECM platform can ease the digitalization of these collaborative CECM business processes and help companies to accomplish a proper transformation of the traditional CECM work area. We also argue that a CECM platform may offer options to mitigate limitations of knowledge required for performing CECM tasks. In particular, this mitigation can be addressed by assistance functions for knowledge-intense assessment tasks and decision tasks in the CECM domain. This articles describes the current status of our ongoing new line of research on the enhancement of CECM information systems by respective assistance functions. The focus is on design considerations for two assistance functions that target i) the CECM task to assess the relevance of new regulatory announcements and ii) the CECM task to determine proper measures to enforce compliance. In particular it is investigated how decision trees can be used to systematically represent the problem space of these tasks and to effectively use the tree to obtain proper task solutions. A two-tier task completion is proposed which first requires domain experts on behalf of the platform providers to prepare a general decision tree. The second step is carried out by the CECM experts of the individual companies. They specialize the general tree to a company-specific tree and then obtain the solution from the tree.

Following a brief description of related work contained in the next Section 2, in Section 3 an overview of the CCPro Cloud Platform is given. Section 4 describes the design considerations for the assistance functions and Section 5 concludes the article with a short summary and an outlook on future research.

[1] Pforzheim University, Pforzheim, Germany, heiko.thimm@hs-pforzheim.de

2. Related Work

In general ICT support for corporate environmental compliance management is not a very active area of research. Only a few articles describe contributions to the research field. A reference model for an environmental management information system for compliance management that makes comprehensive use of business intelligence concepts has been proposed by Freundlieb and Teuteberg [4]. At Stanford University [5] a software infrastructure has been developed that based on semantic technologies offers assistance functions for environmental compliance management tasks. An approach for compliance automation through the use of event monitoring rules has been proposed by a research group of IBM [6]. Also the research of Ly et al. [7] has been focused on runtime monitoring of business processes and among other results lead to a proposal of a framework for compliance monitoring functionalities. Wizards to convey environmental information and to help people to complete environmental management tasks are, for example, described in [8].

The use of decision trees for environmental decision making has been studied by numerous research groups. However, in many of these works a static decision tree is built to a specific reoccurring decision situation. This includes for example decision trees for predicting potential air pollutant emissions [9], decision trees for assessing the effects of exposure to chemical substances [10], and decision trees that model water quality [11]. In this research it is looked at the use of decision trees. In particular it is investigated if decision trees may serve as an effective means to offer system generated guidance and decision support for specific CECM tasks. The decision trees address typical operational tasks of the CECM work practice. Decision tree instances are dynamically orchestrated from scratch based on interactive data from various experts and based on domain knowledge and data about the CECM practice of companies. At the current state of this research a pragmatic orchestration approach is used to generate the trees from the knowledge items. Recently, new theories and models for the generation of sustainability decision trees have been proposed [12]. It can be expected that the various ongoing promising research efforts in the area of Machine Learning will lead to further enhancements of decision tree technology.

3. Conceptual Architecture of the CCPro Platform

With the goal to investigate a Software-as-a-Service (SaaS) approach for environmental compliance management we are developing the Cloud-based System ‘Compliance Center Professional’ (CCPro) [2, 3]. The practical relevance of the research is given by the fact that CCPro’s application functionality is derived from a CECM best practice model [2] that has been obtained in a still ongoing collaboration with industry experts. The following description gives an overview of the conceptual architecture of CCPro that is shown in Figure 1. In particular the major data objects being administered and the central system functionalities are described.

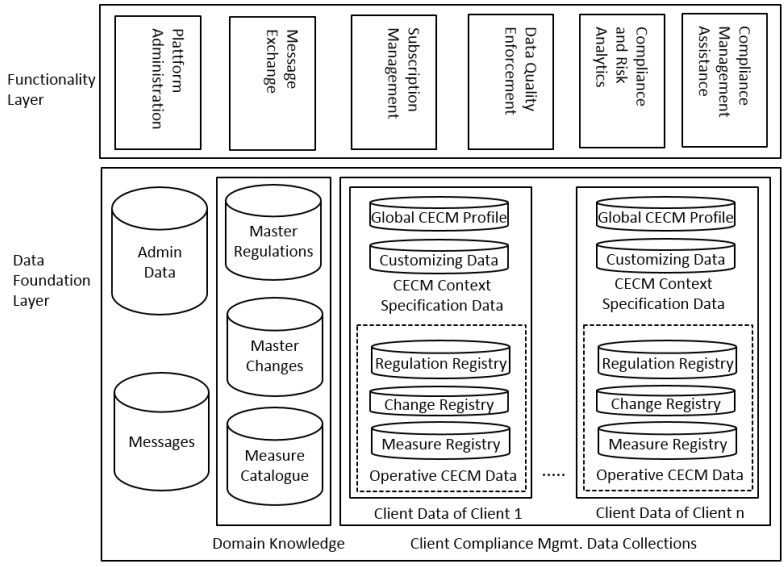


Fig. 1. Conceptual architecture of the Cloud-Based Platform CCPro.

The ‘Data Foundation Layer’ of CCPro contains two data repositories that store administrative data and message data, respectively. Furthermore, the same layer also stores two clusters of data repositories referred by ‘Domain Knowledge’ and ‘Client Compliance Management Data Collections’, respectively. The cluster ‘Domain Knowledge’ consists of three data repositories. The repositories ‘Master Regulations’ and ‘Master Changes’ contain general enhanced digital representations of individual regulations and individual changes i.e. revisions of regulations. By ‘enhanced digital representation’ we mean that the master data objects are enhanced by links to further data objects. These linked objects represent context-specific knowledge in the form of decision trees. As described in the next section assistance is provided to CECM tasks through these instances of decision trees.

The master data objects serve as templates for the creation of company-specific representations of the respective real-world regulations and real-world regulation revisions. These data objects in addition to data derived from the master objects also contain company-specific data such as data of the relevance assessments and data about enforcement measures. The third data repository of the cluster ‘Domain Knowledge’ contains a hierarchically structured catalogue of measures to enforce compliance with environmental regulations ranging from low-effort measures (e.g. provision of information to employees) to measures that require substantial efforts (e.g. enhancements of production facilities).

In general, the data cluster ‘Client Compliance Management Data Collection’ serves as repository for data objects that model real-world objects of companies’ CECM universe. This repository contains an individual collection of client data for every company that makes use of the CCPro platform. Every collection consists of three repositories of operative CECM data and two repositories of data that specify the CECM context of the company. The operative CECM data consists of a set of data from which one can derive the current compliance status of the company. The repositories ‘Regulation Registry’ and ‘Change

Registry' store the client-specific data objects that are instantiated form the above described template data objects. The data cluster 'Measure Registry' contains data about the client's compliance enforcement measures which includes references to the respective regulation or regulation revisions being enforced. Obviously, these references point to the respective data objects in the Regulation Registry and Change Registry, respectively. The repository 'Customizing Data' contains ordinary customizing data as it is usually maintained by business information systems such as data about the organizational structure and the CECM team. The other repository of the 'CECM Context specification Data' referred by 'Global CECM Profile' plays a major role for the assistance functions focused in this article. In principle, the profile data stored in this repository enables the assistance functions to provide context-specific assistance that is specialized to the client's specific CECM situation. The repository content is automatically generated and frequently updated by CCPro from the customizing data and the operative CECM data. The Global CECM Profile consists of a logical map of site-specific CECM profiles for all sites of the company that have been customized in CCPro. Every site-specific CECM profile provides a high-level description of: 1. a list of rule setting institutions that the site has to follow, 2. a list of regulation areas which are relevant for the site due to what is ongoing at the site, 3. data about the local CECM team including the team members' responsibilities and roles, 4. a description of the facilities, products, direct and indirect material being used, and 5. a set of CECM indicators such as the number of relevant regulations, changes, measures, and non-compliance occurrences.

The 'Functionality Layer' of CCPro consists of six functional areas that include general administration functionalities addressed by the area 'Platform Administration' and functionalities for message exchanges addressed by the area 'Message Exchange'. The functional area 'Subscription Management' enables users to subscribe and specialize notification messages that inform about updates of the domain knowledge and about other system events. The functional area 'Data Quality Enforcement' groups functions that are intended to enforce a high data quality such as functions that offer selection boxes and functions that check completeness, consistency, and integrity of data. The functional area 'Compliance and Risk Analytics' offers analytical functions and dashboards specialized to CECM tasks. The assistance functions that are at the center of this article which are intended to support complex knowledge-intense CECM tasks belong to the functional block 'Compliance Management Assistance'.

4. Assistance Functions – Objectives and Design Considerations

Companies are obligated to handle today's constant stream of regulatory announcements by environmental rule setters ranging from regulations of communities up to regulations of supranational organizations such as the United Nations. Best possible efforts are required from the companies to ensure that at all times a state of full compliance with all relevant environmental regulations is maintained at all of the company's locations. In order to meet this requirement complex and knowledge-intense tasks that leverage the power of judgement of experienced CECM experts need to be performed such as:

- a task denoted by Tra that targets to assess the relevance of announced new regulations or new revisions of an earlier announced regulation (or revision),

- a task denoted by Tmd that targets to determine compliance enforcement measures in response to a relevant new regulation or a new revision.

This research targets to enhance the CCPro platform by functions that offer context-specific assistance for the decisions that are implied by these tasks. The functions are intended to promote full use of the power of human judgment without over-challenging CECM resources and without compromising on the risk of non-compliance. An initial approach for the assistance functions was conceptualized that follows a two-tier task completion. A first step is performed by a domain expert without referring to any of the individual user companies of the CCPro platform. This tier one step is intended to narrow down the decision space and to prepare a respective general task completion template. The then following tier two step is completed by the individual CECM experts of the affected companies. In this step the general task completion template is specialized to the company-specific decision situation. Based on the resulting decision space which reflects the relevant company-specific CECM context a proper decision is made and then being applied.

Given this general approach an assistance function denoted by FxRA is conceptualized for the task Tra that is intended to support the domain experts' task. The function through a wizard-based dialog with the domain expert explores the given regulatory situation of a new legal announcement and performs corresponding updates of CCPro's domain knowledge. Then, function FxRa computes a general task completion template in the form of a generic decision tree for relevance assessments of new regulations. A second assistance function denoted by FcRA helps CECM experts to determine a proper company-specific relevance assessment score for the new regulation. Note that the current version of the CCPro Platform supports binary relevance scores, i.e. regulations are either relevant or are not relevant for a company. The guidance of the assistance function FcRA is derived from the generic decision tree, the data repositories of CCPro, and the company-specific context that is queried from the user through a wizard-based dialog.

Also for the task Tmd two assistance functions are conceptualized. The function FxMD is intended to assist the domain experts' task to explore the decision space for compliance enforcement measures and to prepare a corresponding decision tree. The tree is generated on the basis of CCPro's measure catalogue and further context-specific data obtained from the domain expert through a wizard-based dialog. A function denoted by FcMD provides assistance for the next step which is performed by company-level CECM experts. Based on the assistance function the initial decision tree is specialized to a company-specific tree that reflects the particular CECM situation of the company. The specialization makes use of the measure catalogue of the CCPro Platform and also of context-specific data queried through a wizard-based dialog. Note that in principle compliance enforcement measures may consists of a combination of many self-contained measures of different categories. However, the current version of the CCPro platform is limited to "atomic measures" that consist of a single self-contained measure.

The use of decision trees for the targeted assistance functions is motivated by the fact that decision trees [13] have been used as an effective means to structure a problem space and to guide the judgement of options, for example, in selection tasks [14] and classification tasks. In particular, our research targets to explore at a later stage if decision tree learning may serve as an effective computational approach for the intended assistance functions. A corresponding real world data set to experiment with the proposed decision trees can be obtained from an earlier on-premise version of the CCPro system [2, 3]. In particular this

system contains the CECM data of an industry partner which is a mid-size production company in Germany with about 6500 employees at several production sites.

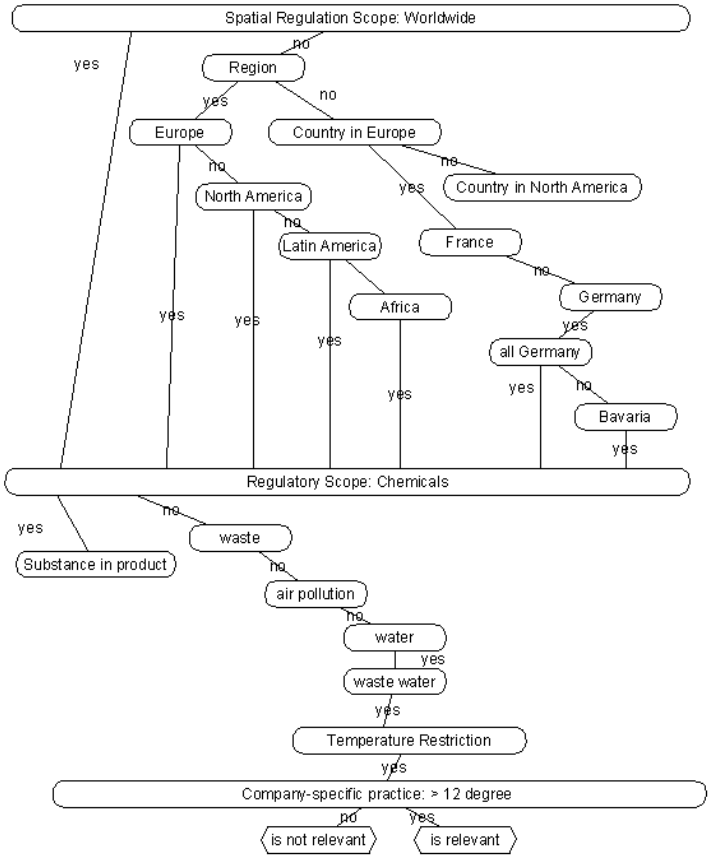


Fig. 2. Fragment of a naive and pseudo kind of RADT tree instance.

For the relevance assessment task the above described assistance functions FxRA and FcRA are proposed. The functions use decision trees to represent, investigate, and determine the context-specific regulatory scope of a particular regulation/revision. The relevance scope is described by a set of conditions that serve as the constituent elements of a respective binary decision tree. Therefore, we refer to these particular trees by 'Relevance Assessment Decision Tree' (RADT). The orchestration process builds a RADT instance by exploring conditions that specify the relevance scope of the regulation in terms of i) the spatial scope, ii) the regulatory scope, and iii) the company-specific CECM circumstances. As one can obtain from the pseudo RADT tree fragment in Figure 2 each of the three areas is addressed by a set of conditions expressed as tree nodes. From the root of the tree to the leaves the conditions are layered as follows. i) The spatial scope of the regulation is addressed by a set of (hierarchical) conditions. The root node condition serves to check if the spatial scope is given by the entire planet Earth. The then following

conditions are based on the common spatial partitioning scheme used to split the planet into regions, countries, states, and communities. ii) The regulatory scope directly follows the spatial dimension. The conditions of the regulatory scope that are captured by the nodes describe which area of the common environmental law areas is addressed by the regulation. Furthermore, every node that specifies a relevant area is followed by a set of further nodes. The further nodes are obtained from a drill-down analysis of the regulation with the goal to explore the specific regulatory item addressed by the regulation. For example, in the tree of Figure 2 the node 'Waste Water' is followed by a node that represents the addressed regulatory item which specifies a temperature restriction for waste water. iii) A set of conditions specialized to both the addressed CECM aspect and the company-specific CECM circumstances constitutes the bottom layer of a RADT tree instance. In principle, the conditions of the regulatory scope are refined and specialized through nodes of the third layer. It is the goal of the specialization to enhance the RADT tree by conditions that consider the CECM context of the given company. The chosen conditions refer to the particular entities affected by the regulation such as products, production processes, materials, equipment, infrastructure, and facilities. In the RADT tree of Figure 2, the temperature restriction is complemented by a specialized company-specific condition that checks the specific temperature range of waste water.

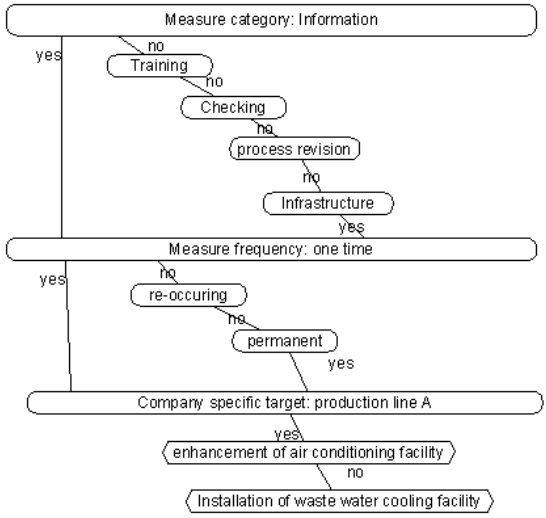


Fig. 3. Fragment of a naive and pseudo kind of MDDT tree instance.

It is the central objective of the assistance functions FxMD and FcMD to provide assistance for the CECM task to determine proper compliance enforcement measures. The functions use decision trees referred by 'Measure Determination Decision Tree' (MDDT) to represent the measure scope, to investigate, and to determine measures to enforce compliance. In particular the measure scope is described in terms of i) the measure category, ii) the frequency type of the measure, and iii) the particular company-specific measure target. As exemplified by the MDDT tree of Figure 3 conditions to describe these three aspects are organized into three layers of tree nodes as follows. i) The nodes of the first layer are intended to identify the proper category of measure. The identification is done based on the measure catalog of CCPro which

among others includes the categories information measure, education/training/instruction measure, checking measure, equipment/infrastructure measure, product revision/redesign measure, and process revision measure. For example, the root node of the sample tree in Figure 3 consists of the condition to check if compliance can be enforced through an information measure. ii) In order to check the measure frequency level a second layer of nodes specifies the frequency categories ‘one time measure’, ‘re-occurring measure’, and ‘permanent measure’. iii) The bottom level of the tree is intended to specify the target object of the measure. In principle, the conditions of the first two levels of nodes are refined and specialized through further nodes in order to tailor a company-specific MDDT tree. Therefore, the conditions of the third layer of nodes refer to company-specific measure targets that can be implied from the actual CECM circumstances. Typical measure targets are products, production processes, materials, elements of the material supply chain, infrastructure components including already existing environmental protection components, machines, and facilities. In the MDDT tree of Figure 3 the permanent infrastructure measure prescribed by the nodes of the first two levels is refined through a corresponding company-specific specialization at the third level. The specialization is expressed by a further node that identifies the infrastructure of Product Line A to be the measure target. The then following two nodes describe options for measures. The leave node represents the installation of a Waste Water Cooling facility to be a proper compliance enforcement measure for this company.

5. Conclusions

The conceptualization of the assistance functions has been finalized so far. A refinement of the function to assist the assessment of the relevance of new regulations (i.e. function FxRA) based on RADT decision tree instances is currently on its way. It is one of the goals of the refinement to explore the state of the art of decision tree theory which may lead to a revised tree structure.

A forthcoming focused implementation will be used for a first evaluation of the decision tree based approach. Top priority is given to an evaluation of the different decision tree scopes. This is due to the fact that the targeted assistance functions require accurate information about the regulatory context of regulations and the CECM context of the companies, too. The evaluation activities will also include end user tests. From these tests we expect insights about the proper decomposition levels of the decision tree scopes and insights for the conceptualization of the wizard-based dialogs.

In general, it is a challenging goal to construct decision trees that offer useful task-oriented and context-specific assistance to users who are responsible for compliance decisions and other CECM tasks. Ideally, decision trees are constructed that ensure that compliance is continuously maintained without overstretching the costs and efforts of enforcement measures. One needs to find an approach that will result decision trees that are well balanced between the risk of non-compliance and the cost of measures.

Therefore, in our future research we will investigate possibilities to use Machine Learning Techniques to obtain well-balanced decision trees that are specialized to the actual CECM situation of the company. It is assumed that the data stored in the Data Foundation Layer of the CCPro platform offers a proper foundation for this approach.

References

[1] Aberdeen Group, “Compliance Management in Environment, Health and Safety: White Paper 6991,” Aberdeen Group, Boston, MA, 2011.

[2] H. Thimm, “IT-Supported Assurance of Environmental Law Compliance in Small and Medium Sized Enterprises,” Int. Journal of Computer and Information Technology, vol. 4, no. 2, 297-, 2015.

[3] H. Thimm, “Towards an Active Assistance and Collaboration Support Platform for Cloud-based Corporate Environmental Compliance Management,” in EnviroInfo: Environmental Informatics - Techniques and Trends: adjunct proc. 32nd edition of the EnviroInfo: Munich, Sept. 2018, H.-J. Bungartz, D. Kranzlmüller, V. Weinberg, J. Weismüller, and V. Wohlgemuth, Eds., Aachen: Shaker Verlag, 2018, pp. 50–55.

[4] M. Freundlieb and F. Teuteberg, “Towards a Reference Model of an Environmental Management Information System for Compliance Management,” in /Berichte aus der Umweltinformatik], Environmental informatics and industrial environmental protection: concepts, methods and tools: Proc. 23rd Int. Conf. Environmental Informatics (EnviroInfo), Sept. 2009, HTW Berlin, Germany, V. Wohlgemuth, B. Page, and K. Voigt, Eds., Aachen: Shaker, 2009, pp. 139–148.

[5] S. L. Kerrigan, “A software infrastructure for regulatory management and compliance assistance,” PhD thesis, Stanford University, 2003.

[6] C. Giblin, S. Müller, and B. Pfitzmann, “From Regulatory Policies to Event Monitoring Rules: Towards Model-Driven Compliance Automation,” IBM Research Rüschlikon, Zürich, Switzerland, 2006.

[7] L. T. Ly, F. M. Maggi, M. Montali, S. Rinderle-Ma, and W. M. P. van der Aalst, “Compliance monitoring in business processes: Functionalities, application, and tool-support,” Information systems, vol. 54, pp. 209–234, 2015, doi: 10.1016/j.is.2015.02.007.

[8] R. Braun et al., “System Architecture and Maintenance of the Ecoradar Web Portal,” in Advanced Information and Knowledge Processing, Environmental Online Communication, A. Scharl, Ed., London: Springer, 2004, pp. 147–160.

[9] D. Birant, “Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models,” J ENV INFORM, vol. 17, no. 1, pp. 46–53, 2011, doi: 10.3808/jei.201100186.

[10] P. Price et al., “A decision tree for assessing effects from exposures to multiple substances,” Environ Sci Eur, vol. 24, no. 1, p. 2212, 2012, doi: 10.1186/2190-4715-24-26.

[11] S. Jaloree, A. Rajput, and S. Gour, “Decision tree approach to build a model for water quality,” Binary Journal of Data Mining & Networking, vol. 4, pp. 25–28, 2014.

[12] K. Doubravsky, A. Kocmanova, and M. Dohnal, “Analysis of Sustainability Decision Trees Generated by Qualitative Models Based on Equationless Heuristics,” Sustainability, vol. 10, no. 2505, pp. 1–18, 2018. [Online]. Available: www.mdpi.com/journal/sustainability

[13] L. Breiman, Classification and regression trees. Belmont, Calif., Pacific Grove, Calif., Pacific Grove, Calif., Monterey, Calif.: Wadsworth International Group; Wadsworth & Brooks/Cole; Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[14] M. G. Vilas Boas, H. G. Santos, L. H. d. C. Merschmann, and G. Vanden Berghe, “Optimal decision trees for the algorithm selection problem: Integer programming based approaches,” Intl. Trans. in Op. Res., vol. 140, no. 1, p. 67, 2019, doi: 10.1111/itor.12724.

Investigation of traffic and air pollution in Thessaloniki, Greece, under ordinary and COVID-19 pandemic conditions

Athanasakis Evangelos[1], Kassandros Theodosios[2], Kostas Karatzas[3]

1. Introduction

Traffic is linked with air pollution especially in urban areas. Even though both parameters follow profiles that have been studied over the years, their investigation provides with new insights when studying their relationships under different conditions. The emergence of the COVID-19 pandemic resulted in lower overall traffic and in the reduction of relevant air pollutants in Thessaloniki, yet this new type of relationship has to be quantitatively evaluated. In this paper we demonstrate how “spot speed” data from IoT sensors, can be processed to capture traffic patterns. Furthermore, we determine the relationship of these data with air pollution levels in Thessaloniki city center, under normal and pandemic conditions and investigate the potential influence of traffic data to the data driven air quality modelling in the area

2. Materials & Methods

2.1. Data availability

Thessaloniki is the second largest city in Greece and the financial, educational and cultural center of the Macedonian region in the north part of the country. The city is densely populated and traffic is among the main contributors of the relatively high air pollution levels in the area. The city center is characterized by a mixed, commercial and residential use, has a high number of narrow streets and accumulates major parts of the everyday human activity of the city. Traffic data are made available via the "TrafficThess" portal (http://www.mobithess.gr/), which is part of a larger project called "Intelligent Urban Mobility Management System". Data are based on the mean travel speed of taxies used as traffic proxies. On the other hand air pollution data originate from the official air quality monitoring station operated in the city center (Agia Sofia), via the web site of the European Environment Agency (EEA). For traffic, a dataset with 9 roads, within a 100m radius around “Agias-Sofias” and for a time period of almost 5 years (2015-2020) is engineered. This dataset includes data for the COVID-19 lockdown – related months as well, while for pollutants those data were downloaded separately from EEA’s portal.

2.2. Traffic parameters

[1] Aristotle University of Thessaloniki, Thessaloniki Greece, vaggelis.atha1993@gmail.com
[2] Aristotle University of Thessaloniki, Thessaloniki Greece, tkassand@physics.auth.gr
[3] Aristotle University of Thessaloniki, Thessaloniki Greece, kkara@auth.gr

Even though it is known that vehicle density (vehicles/km) is expected to produce more accurate results for emission estimations [1], a transition to this kind of data requires either the use of inaccurate deterministic models [2], or the extra availability of traffic flow data (which are rare), to train and validate neural networks that estimate the last through mean speed inputs, so that density can be subsequently estimated through the speed-density equilibrium [3], [4].

Nevertheless, it is possible to create a congestion parameter by using the hourly average of the "spot speeds" in a road, as well as, the "free flow speed" of the same road, this parameter is demonstrated by equation 1 and it is called traffic performance index (tpi) [5], where u_f is the free flow speed which is unique for each road and u are the hourly average "spot speed" observations for the road. The tpi scales from 0 (free flow) to 1 (full congestion).

$$tpi = \frac{u_f - u}{u_f} \quad (1)$$

In order to create a single tpi variable for an area (rather than using each road as a separate variable), we average the tpi of all available roads in that area. As a condition, we demand that a road has a mean correlation of 0.25 or higher with all the other roads, before including it in the average. This way we check if a road's traffic conditions are related with those in the rest of the area, or if it is for example a small segment close to intersections or traffic lights where traffic conditions cannot effectively be captured by speeds. Overall, the accuracy in the estimation of the actual traffic flow conditions via the tpi is greatly affected by the road segments included in the calculations as well as by the traffic conditions in these areas (the denser the traffic, the higher the accuracy)

2.3. Handling outliers

For the case of pollutant data the sensors are exposed to the outer environment so outlier detection and removal is necessary. The common IQR (inter-quartile range) method [6] is used to achieve the above and in python, numpy[4] can help automate this pre-processing step so that it works fast on large datasets. In the case of speed data, we have concluded that no extreme values are present. Removing outliers would cause the loss of valuable, close to free-flow-conditions information. The only transformation necessary in this case, is to scale down speeds which are larger than the free flow speed, since those indicate free flow conditions anyway.

2.4. Spotting seasonalities

The best way to infer the periods in time-series data is through a Fourier transform (FT) and specifically a discrete one (since we are interested in discrete data). Fourier transform, aims at expressing a time-series through a sum of sines and cosines. The Fourier transform allows the transition from a time representation to a frequency one, thus making it possible to determine the basic periods in a signal. Python's scipy[5]

[4] https://numpy.org/
[5] https://www.scipy.org/

library**,** an ecosystem of open-source software for mathematics, science, and engineering, was used to perform Fourier transforms.

2.5. Handling missing values

For performing Fourier transforms, we use mean imputation, which equally affects the power spectrum and therefore does not significantly alter the seasonalities. For the rest of the analysis, in order to boost the significance of our results, even though the methods can work with missing values, we use hot deck imputation at least for the case of traffic data where the missing values don't form large chunks, and different roads are actually expected to behave similarly. For the case of pollutants the above don't apply and there is no obvious solution to imply missing values.

The hot deck imputation method that was used, is k-NN [7]. The Euclidean distance is used as a similarity metric and $k = \sqrt{N}$ neighbors are selected [8]. The mean of those k neighbors is the imputation value. The different features that act as neighbors in the case of traffic data are the tpi values for the rest of the roads around an area. In python it is easy to implement this kind of imputation through the sci-kit[6] learn library, which is a collection of basic machine learning and data pre-processing algorithms.

3. Results

The Fourier transforms provide the following insights:

- A seasonality of 1 day followed by one of 12 hours is characteristic in traffic data
- A seasonality of 12 hours followed by one of 1 day is characteristic in NO_2
- A seasonality of 12 hours is also apparent in $PM_{2.5}$
- In all cases, weaker seasonalities close to 1 week are observed

The information above emphasizes the importance of studying the average diurnal behavior of pollutants and traffic. The pandas[7] library, provides quick and effective modules to manipulate data, while matplotlib[8] and seaborn[9] (which is built on top of matplotlib) can produce interesting graphs. Figures 1 through 3 provide some visualizations of traffic patterns in time. It is evident that during weekends traffic (in terms of tpi) is lower. The same applies to summer months and especially for August, while during fall and winter months traffic is increased. Concerning everyday patterns, Tuesday and Thursday afternoons are the most congested periods in terms of traffic.

[6] https://scikit-learn.org/stable/
[7] https://pandas.pydata.org/
[8] https://matplotlib.org/
[9] https://seaborn.pydata.org/

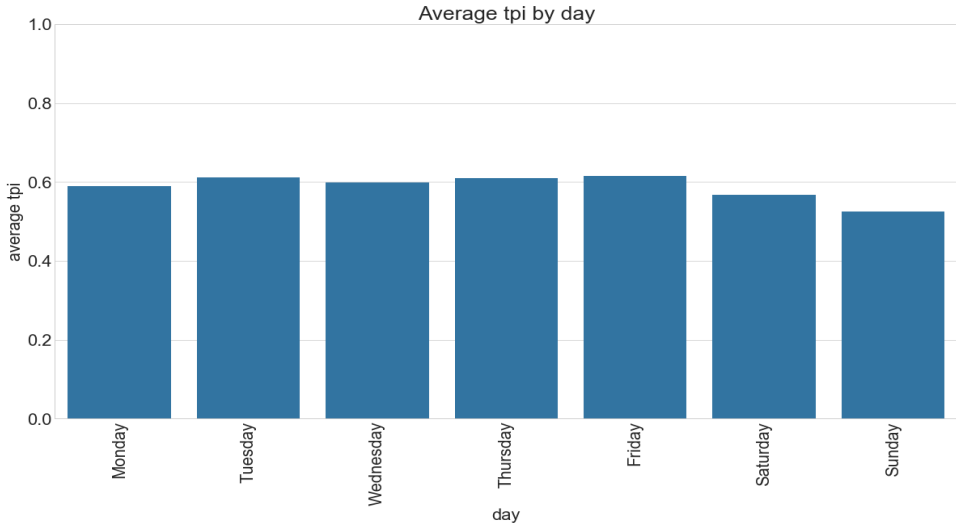


Fig 1. Average traffic by day.

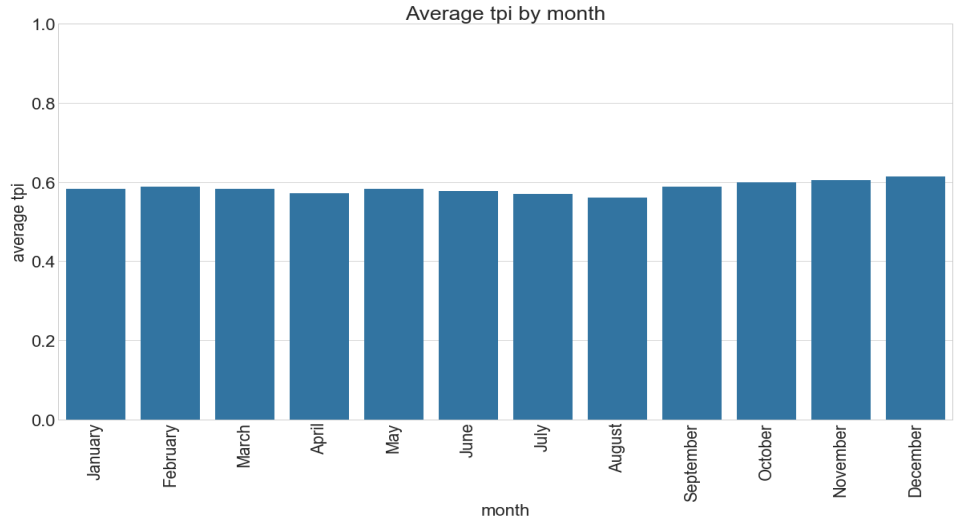


Fig 2. Average traffic by month.

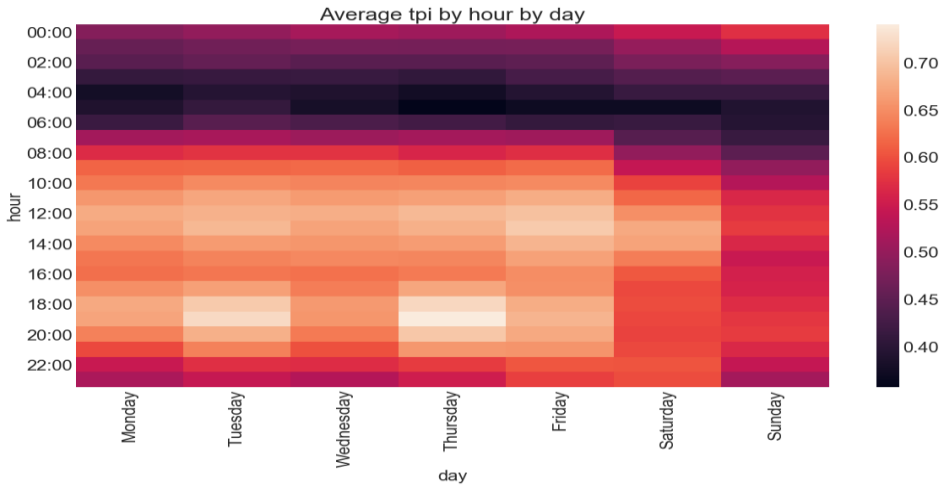


Fig 3. Average traffic by hour and by day.

Figure 4 present with the normalized, average diurnal profiles of NO_2, $PM_{2.5}$ and tpi. These pollutants typically display two peaks, one in the morning and one in the evening (as a result of both emissions and meteorological conditions), while traffic mainly peaks around noon and afternoon. The NO_2 profile shows a lagged behavior compared to the $PM_{2.5}$ one. The fact that tpi peaks when pollutants decrease in afternoon, or that pollutants peak at night when tpi decreases, is an indication that meteorological phenomena mainly drive pollutant concentrations. Nevertheless, a simultaneous increase in tpi and $PM_{2.5}$ is observed in early morning hours

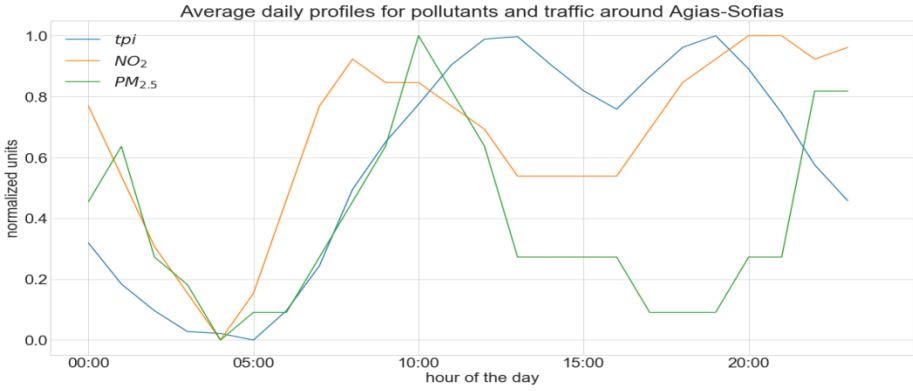


Fig. 4 Average diurnal profiles of tpi and pollutants.

Figure 5 shows the average diurnal profiles of tpi, during the COVID-19 lockdown period, for 2020 (23 March – 3 May) and for the same period for previous years for which data are available, so that a comparison can be made. The decrease in traffic is obvious, though it must be noted that results in the tpi must be interpreted with caution in this case. Vehicle traffic decreased during lockdown, so the samples used to generate the traffic data are much smaller, thus less representative. Even after averaging all roads, approximately 33% of the desired data were missing and most of them correspond to early, after-midnight hours. We used the k-NN imputation to calculate missing values, which might have caused the tpi to be overestimated. The profile's behavior shows a slight change, since the daily peak now seems to decrease faster in the evening. Figure 6, shows a graph similar to Figure 5 but now for NO_2. The decrease as well as the changed behavior in concentration values, is obvious. While the two daily peaks used to be almost equal in magnitude, the lockdown has caused the night peak to be much smaller. Finally, Figure 7 shows the same profiles for the case of $PM_{2.5}$. The strongest morning peak has now become weaker than the one at night. This, along with the findings reported by Figure 4, could imply that the $PM_{2.5}$ concentrations increasing in morning-noon hours are directly related with traffic, while the night increases might be due to other emission sources and might reflect influences of meteorology. On average, we report the following decreases due to Covid-19 lockdown, in comparison with the same period for previous years:

- A 47.14% decrease in NO_2 concentration levels along with a 20.68% decrease in traffic
- A 30.37% decrease in $PM_{2.5}$ concentration levels along with a 21.06% decrease in traffic

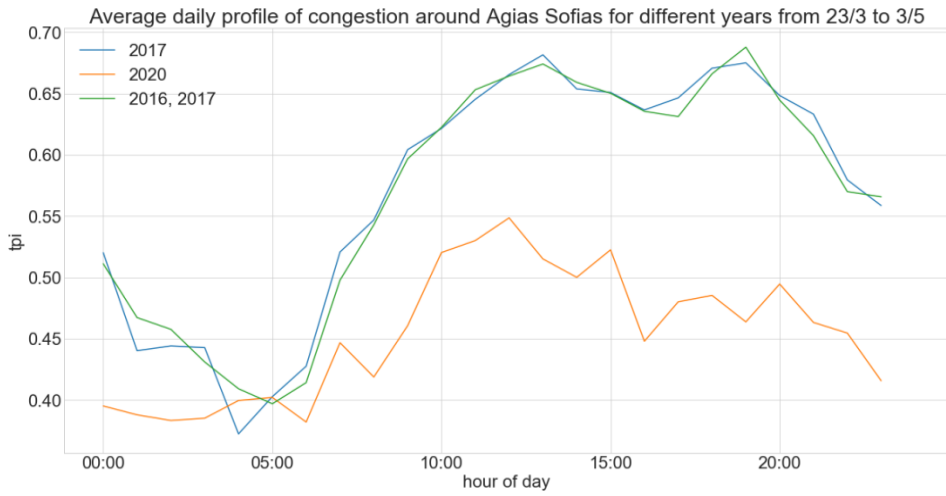


Fig 5. Average traffic diurnal profiles for Covid-19 lockdown period.

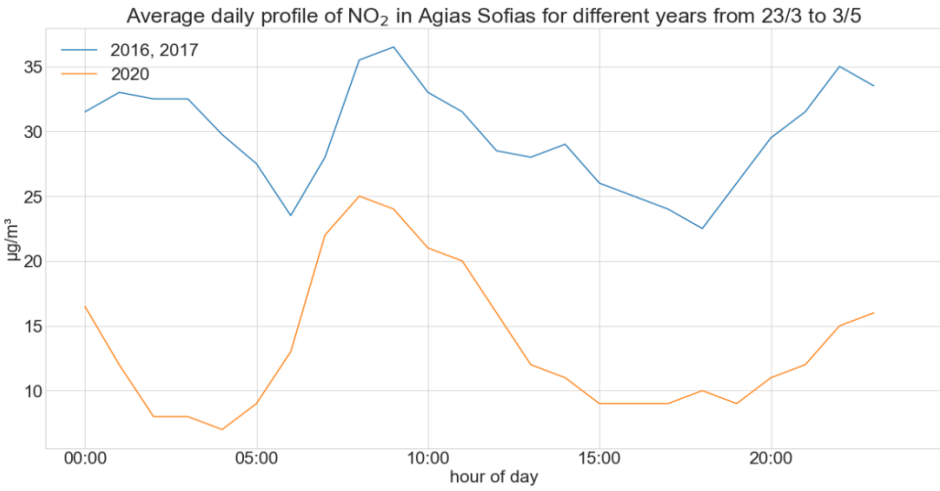


Fig 6. Average NO_2 diurnal profiles, for Covid-19 lockdown period.

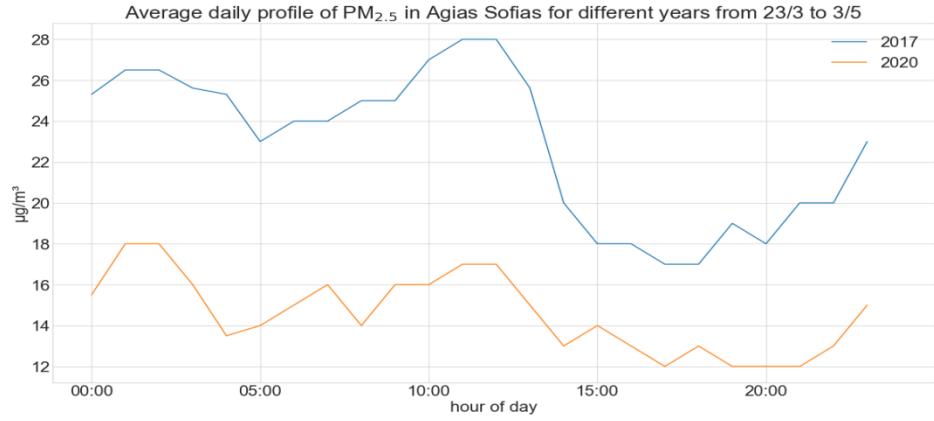


Fig 7. Average $PM_{2.5}$ diurnal profiles, for Covid-19 lockdown period.

Finally, Figure 8 shows the correlations of tpi and pollutants, for different hours during the day. Even though the overall correlations of tpi with pollutants are weak (0.28, 0.26, 0.13 for CO, NO_2 and $PM_{2.5}$ respectively), the relationships seem to be moderate during the day and too weak, or even negative at night. This supports the intuitively expected result that even for important emission sources of air-pollutants (like traffic) it is not very probable to observe strong correlations, because the atmosphere is a dynamic system where meteorological phenomena are expected to greatly affect the concentrations of pollutants. Also CO and $PM_{2.5,}$ seem to have the strongest relationships in morning and noon while NO_2 in afternoon. This graph indicates the necessity of including tpi features in data driven air quality forecasting models with hourly dependent approaches.

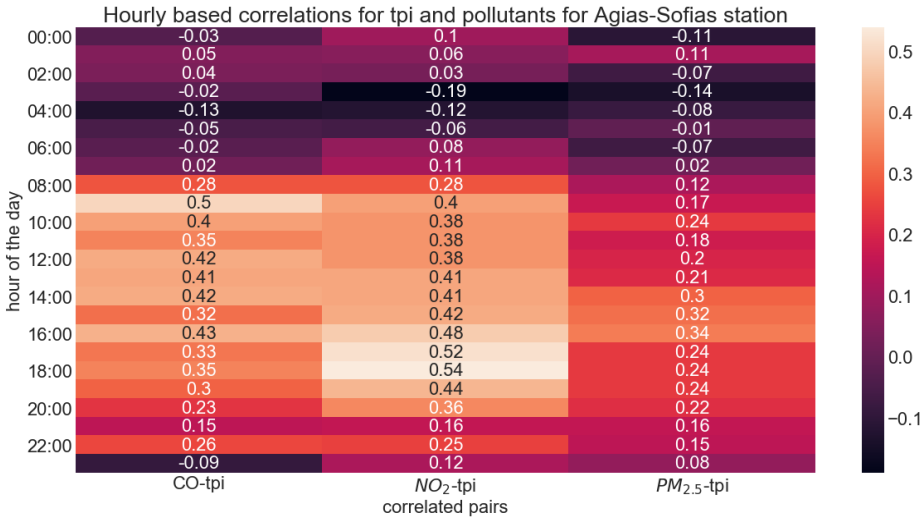


Fig 8. Traffic-pollutant correlations by hour.

4. Discussion

Even though we demonstrate that a traffic congestion index like tpi can reflect traffic patterns effectively, the overall correlations with pollutants, are weak. Nevertheless, we demonstrate that the relationship can vary throughout the day reaching moderate correlations, so tpi features could potentially contribute to air quality forecasting models if that variation is considered. Providing continuous access to the traffic data through a database and creating an application programming interface (API), can open the road for future works to study potential transitions to traffic parameters like flow or vehicle density which are expected to be more related with pollutant concentrations. The ideal dimensions of a representative grid-cell around an area need to be studied as well, along with potential dimensionality reduction methods that might perform better than simply averaging the traffic conditions of all the roads within the cell. On the other hand, independently of the potential improvements mentioned above, average diurnal profiles of tpi and pollutants can provide valuable insights.

Moreover we have demonstrated the altered average diurnal behavior in tpi, PM2.5 and NO2, due to the COVID-19 pandemic lockdown and doing so for other pollutants could help field experts draw valuable conclusions as to how traffic impacts air-pollution overall.

Acknowledgments

The authors acknowledge “CERTH-HIT” for providing free access to traffic data through the “TrafficThess” portal.

References

[1] Ryan P. H. and LeMasters G. K. *A review of land-use regression models for characterizing intraurban air pollution exposure*. In: Inhal Toxicol vol. 19 Suppl 1 (2007), pp. 127–133.

[2] Yu C. et al. *"Speed-Density Model of Interrupted Traffic Flow Based on Coil Data".* In: Mobile Information Systems 2016 (2016), p. 12.

[3] Florio L. and Mussone L. *"Neural-Network Models for Classification and Forecasting of Freeway Traffic Flow Stability."* In: Control Engineering Practice 4.2 (1996), pp. 153–164.

[4] Smith B. L. and Demetsky M. J. *"Short-term Traffic Flow Prediction: Neural Network Approach."* In: Transportation Research Board 1453 (1996), pp. 98–104.

[5] Hongsuk Yi, HeeJin Jung and Sanghoon Bae, *"Deep Neural Networks for traffic flow prediction,"* 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), Jeju, 2017, pp. 328-331, doi: 10.1109/BIGCOMP.2017.7881687.

[6] Upton, Graham; Cook, Ian (1996). *Understanding Statistics*. Oxford University Press. p. 55.

[7] Batista G.E.A.P.A. and Monard M.C. *"A study of k-nearest neighbour as a modelbased method to treat missing data"*. In: Proceedings of the 3rd Argentine Symposium on Artificial Intelligence. Vol. 30. Buenos Aires, Argentine, 2001, pp. 1–9.

[8] Duda R.O. and Hart P.E. *Pattern Classification and Scene Analysis*. NY: John Wiley and Sons, 1973.

Machine learning methods for approximating the temperature of exterior walls using thermal images and colour images of building facades

Klaus Schlender[1], Malte Riechmann[2], Jendrik Müller[3], Grit Behrens[4]

1. Introduction

Houses lose energy in the form of heat through their outer walls. The amount of heat lost depends on the structure of the wall. New buildings in Germany already have a high energy standard according to EnEV 2016 [1] and are built according to the new EU building regulations [2]. But the lots of old houses with low heat insulation lose energy. The building sector causes about 30% of all CO2 emissions and heat insulation is of great importance when considering climate protection goals [3].

Modernization of older houses is expensive and should be considered carefully. In the ENVIRON project "Environ - Entwicklung und Evaluation einer Intervention zur Vermeidung von durch energetische Sanierung ausgelösten Rebound-Effekten" (BMBF[5] joint research project, Funding number: 01UT1703A), buildings are considered in cooperation with "Reallabor Bielefeld Sennestadt" [4]. Apartments in the buildings were measured using a SmartMonitoring system. Data from the rooms of buildings before and after the refurbishment were collected [5-6] and thermal images of the exterior wall were also taken to record its thermal radiation.

Thermography is becoming more and more important in the renovation process. [7] This makes it possible, for example, to detect thermal bridges, observe heat transfer, or calculate heat losses with physical formulas. Nowadays, thermal images are often used for the evaluation and detection of thermal bridges in house walls. Besides thermal bridges, thermal areas and their heat radiation are also considered by experts.

These thermal images are easy to capture, but need to be analyzed by an expert. For this reason, the focus of this work is to enable an automation of thermal image processing. This can help experts to analyze and compare buildings before and after their refurbishment with each other. A method will be developed in further project that offers a faster and simple way to calculate the thermal energy radiated by outside walls. This requires in the first step the automated removal of noise objects in the termal images, that are not part of the building, such as the sky, or those that are causing noise on the thermal image, such as windows.

However, the work in progress in this paper can contribute to a faster evaluation of potential energy losses in buildings that require refurbishment and help to support the decision-making process for or against a modernization process in the future. Thereby it will also do its part to help reducing overall C02 emissions as well as benefiting the environment.

[1] FH-Bielefeld, University of Applied Sciences, Minden Germany, klaus.schlender@fh-bielefeld.de
[2] FH-Bielefeld, University of Applied Sciences, Minden Germany, malte.riechmann@fh-bielefeld.de
[3] FH-Bielefeld, University of Applied Sciences, Minden Germany, jendrik.mueller@fh-bielefeld.de
[4] FH-Bielefeld, University of Applied Sciences, Minden Germany, grit.behrens@fh-bielefeld.de
[5] Federal Ministry of Education and Research, via: https://www.bmbf.de/en/index.html

2. Methods

2.1. Thermal images and their information values

Thermal imaging technology is based is the so-called temperature radiation. Temperature is basically the disordered movement of particles [8]. In order to visualize and evaluate the temperature radiation via a camera, it is necessary to understand its properties. First of all, temperature radiation, like any other electromagnetic radiation, is a source of energy. This is used by the infrared camera to visualize the temperature.

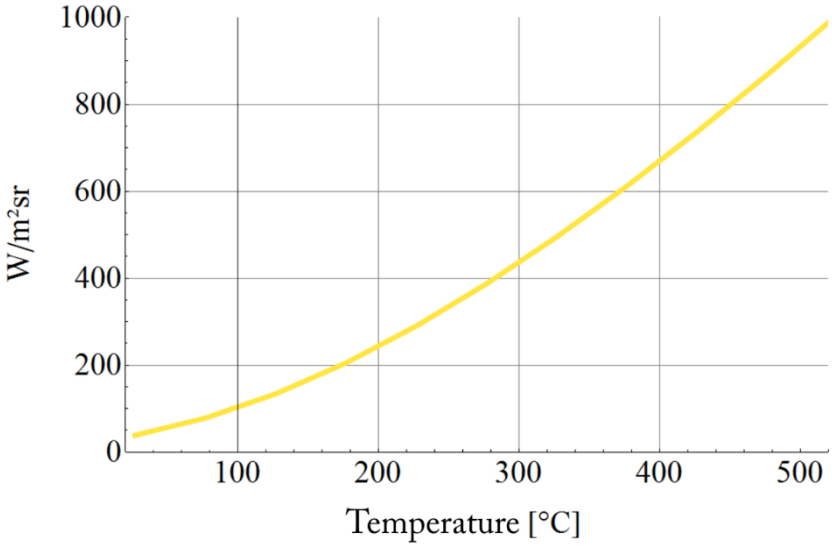


Fig. 1. The relationship between temperature and the integral of Planck's radiation law [9].

To give a brief comparison and not to go into too much detail. An infrared camera differs in many aspects from a usual digital camera. The first is the wavelength range in which the camera is operating. The digital camera reproduces the spectrum visible to humans, while the infrared camera records a section in the infrared range [10]. The usual range is 1 - 12 µm. In addition, the infrared camera does not capture colors, so it does not differentiate between the various wavelengths in a range. The Planck's law of radiation [11] is used regarding what the camera records. Warmer surfaces have a larger radiation area and appear as brighter objects. The integral of the curve can be mapped against temperature to obtain the corresponding temperature for each brightness. An example curve for this is given in Figure 1.

Fig. 2. Thermal image of a building from the city Bielefeld-Sennestadt.

Thermal images show the different temperatures of the environment. Simply put, each pixel holds information about the temperature at the corresponding position, but the temperature information is converted to visualize it as an image. Basically, thermal images could be treated as grayscale images, and not color images as in Figure 2, because each pixel encodes exactly one temperature value. These values are normalized according to equation (1), based on the minimum and maximum temperature seen in the image.

- $I(x,y) = 255 * \frac{(T(x,y) - Tmin)}{Tmax}$ (1)

In this function, I is the resulting grayscale image having I(x, y) as the pixel at position (x;y). Similarly, T(x, y) is the measured temperature at the position. Tmin and Tmax are the minimum and maximum measured temperature in the image. The resulting value ranges between 0 and 1. Since images usually use a byte representation for one pixel, this value is multiplied by 255. The reason why the normalization is done, is to increase the contrast between the different temperatures. This makes it easier to detect slight temperature differences. Furthermore, this way there is no problem when dealing with negative temperatures, which can occur, for example, when the temperature is given in Celsius. However, grayscale images are still more difficult for humans to interpret. For this reason, in visualization of an IR-image a color is assigned to each temperature value. This assignment is described by a color palette. An example of the used color palette is shown on the right side of the picture in Figure 2.

In addition to the pixel values, the image contains some metadata to provide additional information. The metadata is stored in EXIF format and contains basic information that is used in further image processing. To mention some of the more important metadata, the resolution and color space information is useful when normalizing the images in the pre-processing stage. Other metadata like creation date are

used to pull measurement data from the measurement system from the same time stamp from its corresponding room. Weather data is also available and is read out to be used in the further processing.

2.2. Preprocessing and supervised labelling of thermal images

Before the thermal images are used to train a Convolutional Neural Network (CNN), they must be pre-processed. In this section, various image preprocessing methods are shown.

The first step is to standardize the color spaces of thermal images, because different infrared cameras store their data in different color spaces. The respective color space can be taken from the metadata of the image. The infrared camera used for the project has adjusted and applied standard RGB color space. To normalize the images, colors stored by infrared camera in the images must be exported as highest and lowest temperature values. Additional data, like the color palette being used, is also retrieved and can be accessed using metadata of images. Afterwards colors can be converted into temperatures. This allows the image content to be converted into absolute temperatures by changing equation (1) from chapter 2.1.

In the images taken for our project thermal images always show some base heat level. The sky, on the other hand, is always much colder and mostly shown in black tones. This can therefore be filtered out by eliminating all areas whose temperature does not reach a certain threshold value. The formula used for this can be seen in equation (2).

- $$I(x,y) = \begin{cases} I(x,y), & I(x,y) > t \\ 0, & otherwise \end{cases} \quad (2)$$

In addition to thermal images, there are also available simultaneously recorded colour images, which were taken at the same position. The resolution had to be scaled uniformly to allow them to be usable later on. For further processing in CNN training, all images were prepared and areas were labelled (Figure 3) in order to train object detection and also to be able to evaluate the training process.

Fig. 3. Example of a supervised annotated image.

To create labels as masks on the images a free available tool on the domain supervise.ly was used. Masks were drawn on the original image and saved as JSON files for further use. The labeling was done on the colour images, because edges are better visible on them. Since colour images and thermal images are overlapping, the created labels can be used for both of them.

After images are labeled, the CNN network can be taught what to recognize. In our case, the goal is to detect windows in order to remove them as automatically detected masks. In this way they will not affect the calculation of the approximate average temperature of the outer wall afterwards.

To enlarge the data set of usable images two methods of data augmentation were applied. The first method is random crop. This involves cutting out a random area of the image and scaling it to uniform size. The same is done with the mask to train to the correct label. The second method is Random Flip Left-Right. In this method, the input and the mask are mirrored along the vertical axis. Further augmentation methods were not used because some methods can result in loss of information when using them on thermal images. In order to keep all approaches comparable (various experiments from chapter 2.3), we have chosen to limit to these two augmentation methods.

2.3. Automatic window detection during experiments

To detect windows, 350 images were randomly split 60:40 into training and test data set. The labels were created supervised as described in the previous chapter.

Both the training data and the unseen test data contain images, each having a mask and a bounding box as label for each window.

The architecture of CNN shown in Figure 4 was used for the window detection. It is based on the U-Net architecture [13]. The first phase of this architecture is used for feature generation by the backbone. A pre-trained MobileNetV2 was used for this purpose [14]. The generated feature maps are saved from each convolutional step. After the last convolution the up sample phase follows. For each convolution of the backbone there is one up sample step. Each up sample receives as input the output from the previous step and the output from the corresponding convolutional step. For this reason, the generated feature maps were saved after each convolution. At the end the network outputs an image with two channels in the same size as the source image. These two channels are the One-Hot-Encoded class-labels of each pixel. The classes that are learned are window and non-window and can be used to detect and locate unseen images afterwards.

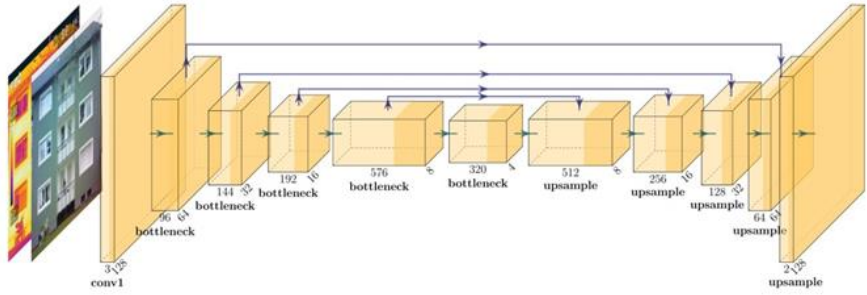


Fig. 4. Used architecture to generate mask areas to detect windows.

Since we use MobileNetV2 which is only designed for three channel input images, but in our work we also use one and four channel images, an additional layer was placed in advance of MobileNetV2. This layer only helps to bring the input image onto three channels in order to be able to feed it into the backbone.

Three experiments were conducted to train window recognition. The first training was based on thermal images only, the second training was based on colour images only and the third training combined both infrared and colour images. Figure 5 shows the loss function of the three experiments on window detection. The blue line shows the run of the loss on the validation data and the red line shows the run on the training data.

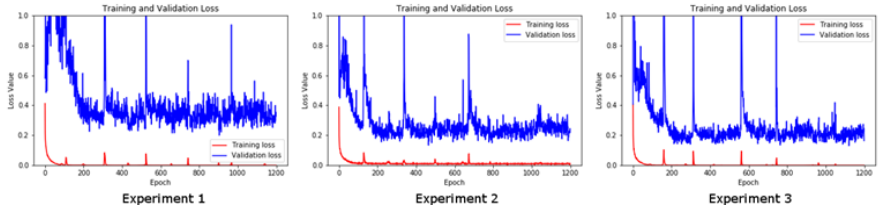


Fig. 5. The three experiments on window detection.

In the first attempt only thermal images were used as input for the network. The network was trained for 1200 epochs and had a final validation loss of 0.33. The process is shown in Figure 5 as experiment 1.

There you can see some large fluctuations in the loss, which indicate that the network has problems to recognize a window based only on the thermal image. Figure 6 shows an example result for this experiment. You can see that the result is very good in general, because all windows were detected. But the edges of the windows are not in focus and there are small holes in the window regions (Predicted Mask). There are also more often some pixels that are recognized as a window region but do not belong to one.

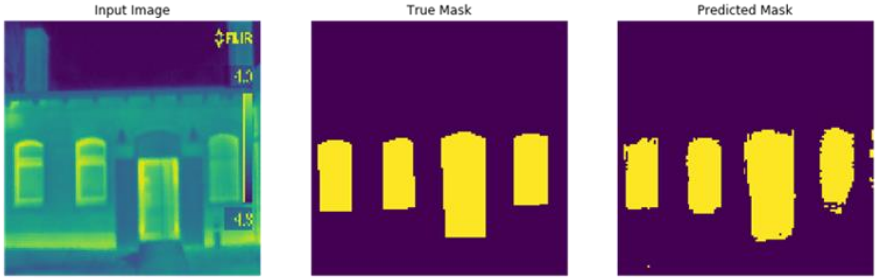


Fig. 6. Example result of the first experiment.

In this second experiment only colour images were used as input. Again, the net was trained for 1200 epochs, with a validation loss of 0.22. In Figure 5 on Experiment 2 the evolution of loss is shown. There can be seen, that the loss is generally lower compared to experiment 1 and that the loss is less fluctuating. This means, that it is easier for the network to train the given data samples. The same is reflected in the qualitative analysis. An example result can be seen in Figure 7. Compared to the previous example, there are much sharper edges and the objects predicted fit much better to the expected forms.

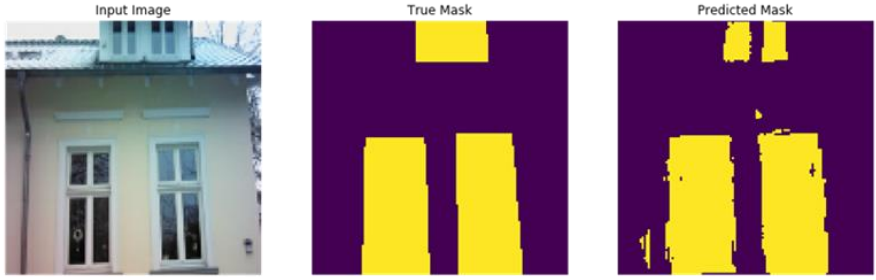


Fig. 7. Example result of the second experiment.

In the third experiment, both colour images and thermal images were used. In this experiment, a four-channel combined image was created as input. For this purpose, the thermal image was included as an additional color channel. As in both previous experiments, the network was trained for 1200 epochs. The loss function can be seen in Figure 5 (Experiment 3). At the end of the training, the loss was 0.16. This is the lowest and best loss achieved in the detection experiments. Also, the evolution shows less fluctuations and a flatter course. The qualitative analysis confirms these impressions. Even for challenging tasks on images with acute angles, as shown in Figure 8, the network can deliver well-defined masks with less holes than was the case in the other experiments.

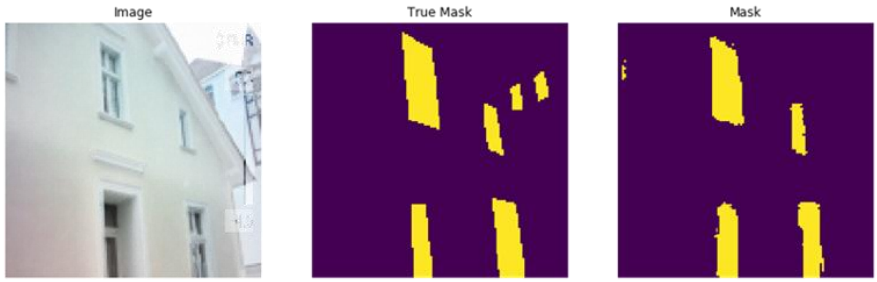


Fig. 8. Example result of the third experiment.

With the trained CNN from experiment 3, window areas can be recognized effectively and removed from a given image to avoid potential negative noise. In this way, the windows of a selected region of the outside wall ,which belongs to an apartment, are removed using automatically generated masks, in order to be able to calculate the approximated average temperature of the outside wall across its pixels in following process.

The approximately calculated average temperature value of the outer wall can then be used together with the temperature of the inside wall as measured by the measuring system, by thermography experts. They will then be able to calculate the heating energy losses of the outside wall area at a certain time or over a heating season.

3. Results

After already available values for comparison were extracted from measurement data of the measurement system and from the metadata of the images, further unknown variables could be successfully computed from the thermal images. In this case, the temperature of the outside wall was taken from the average of the thermal image's temperature values. The unwanted areas such as windows could be isolated using machine learning methods with image segmentation using object detection before the calculation was performed.

Based on the acquired, measured and calculated data, an automated and simple image processing can be performed. The data can later be used by thermography experts for approximation of energy loss through the outer walls. The experts can then better decide whether a refurbishment is relevant or not. However, it is also necessary to consider under which environmental conditions the images were taken as well. For example, if the image was captured under warmer outside temperatures, the need for refurbishment may be less than it would have otherwise seemed in colder regions.

However, a more accurate conclusion about the relevance of a renovation requires a larger amount of data than is available within this research project. If more data from buildings on different weather conditions could be included in the future, it would be possible to estimate energy losses.

4. Discussion

The research shows that with the help of artificial intelligence it is possible to achieve adjusted results for thermal heat in the average value and can be used in the future to calculate thermal energy losses. The use

of image segmentation and therefore the removal of noise objects using object detection results in a more accurate approximation than if just a simple average value is calculated over all pixels.

The existing system can be further improved and extended. By creating additional labels for different objects, even more potential noise can be filtered out, which also improves the estimation accuracy. In further work, the thermal energy losses for other objects could be included in the algorithm. For example, heat losses from doors, walls and windows might be integrated separately into the calculation. This may improve the estimation of energy loss even more.

Acknowledgments

The authors thank the Bundesministerium für Bildung und Forschung and FONA Sozial-ökologische Forschung BMBF for their financial support of the project "Environ - Entwicklung und Evaluation einer Intervention zur Vermeidung von durch energetische Sanierung ausgelösten Rebound-Effekten", Funding number: 01UT1703A.

References

[1] EnEV: Neubau Wohngebäude ab 2016. https://enev-online.com/enev_praxishilfen/vergleich_enev_2016_enev_2014_neubau_wohnbau_15.04.06.html (2016). Accessed 10 APR 2020

[2] Implementing the Energy Performance of Buildings Directive (EPBD). Lisbon, September 2015, ISBN 9789728646325, Via: http://www.epbd-ca.eu/outcomes/2011-2015/CA3-BOOK-2016-A-web.pdf (2016). Accessed 10 APR 2020

[3] Umwelt Bundesamt: Treibhausgasminderungsziele Deutschlands. https://www.umweltbundesamt.de/daten/klima/treibhausgasminderungsziele-deutschlands (2020). Accessed 19 FEB 2020

[4] Sennestadt GmbH: Stadtlabor Sennestadt. https://www.sennestadt-gmbh.de/stadtlabor-sennestadt.html (2020). Accessed 19 FEB 2020

[5] Behrens, G., Hamelmann, F., Thiel, C., Försterling, T., Weicht, J., Fehring, F., Schlender, K., Dreimann, R. (2017). Smart measuring system of air quality accompanying a renovation process of apartment buildings. EnviroInfo 2017, Luxemburg, Springer Nature 2018, imprint in “From Science to Society – new Trends in Environmental Informatics” ISBN 978-3-319-65686-1

[6] Behrens, G., Schlender, K., and Fehring, F. (2018). Data mining methods of healthy indoor climate coefficients for comfortable well-being. Environmental Protection and Natural Resources; The Journal of Environmental Protection - National Research institute. 29, 3, 7-12, Available From: De Gruyter - Sciendo. Via: https://content.sciendo.com/view/journals/oszn/29/3/article-p7.xml Accessed 19 FEB 2020

[7] Pleşu, Raluca & Teodoriu, Gabriel & Taranu, George. (2012). INFRARED THERMOGRAPHY APPLICATIONS FOR BUILDING INVESTIGATION. Bulletin of the Polytechnic Institute of Iasi - Construction & Architecture Section. 62. 157-168.

[8] Schroedel, editor.Physik: Ausgabe Dorn-Bader. Gymnasium Sek II. Ge-samtband. Schroedel, 2001

[9] Johannes Horak. Von schwarzkörper im bereich 8-12 mikrome-ter bereich abgegebene lichtintensität pro raumwinkel, https://www.timaios.org/wp-content/uploads/2014/11/planck_integriert_detektierte_intensitaet.png Accessed 10 May 2020

[10] Johannes Horak. Wie funktioniert eine infrarotkamera? temperatur-bestimmung. https://www.timaios.org/2014/11/27/wie-funktioniert-eine-infrarotkamera-temperaturbestimmung/ Accessed 5 Jan 2020

[11] Johannes Horak. Das planck'sche strahlungsgesetzt ausgewertet für körperunterschiedlicher temperatur. https://www.timaios.org/wp-content/uploads/2014/11/planck1.png Accessed 10 May 2020

[12] studyflix. Stationäre wärmeleitung. https://studyflix.de/ingenieurwissenschaften/stationare-warmeleitung-473 Accessed 10 May 2020

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation (2015)

[14] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks (2018)

Induction of a fuzzy decision tree for optimizing air quality data modeling

Aristotelis Karagiannis[1], Kostas Karatzas[2]

1. Introduction

Air pollution (AP) is one of the most significant global scale problems of our time. This phenomenon has serious impacts on public health, deteriorates the quality of human life and contributes to the degradation of environmental quality. The development of reliable models for predicting AP episodes is nowadays feasible, due to the vast volume of available data, which can be exploited by machine learning algorithms. A common drawback of this kind of algorithms is their inability to provide a comprehensible interpretation of the principles that define the operations of the system being studied [1]. This problem can be tackled by employing fuzzy logic rather than classic logic, through the theory of fuzzy sets.

In the current paper we develop an algorithm for the induction of a fuzzy decision tree, which is then materialized using the Python programming language. The fuzzy decision tree is trained according to a dataset, which contains air quality measurements from the greater Thessaloniki area in the Macedonia region of northern Greece, and it is compared with other data-driven models, which are based on classical logic. A unique characteristic of every fuzzy model is its ability to describe the system through a set of fuzzy rules, where the conditions and the consequences are stated by expressions of the physical language [2]. The results show that the fuzzy decision tree, even in this prototype form, excels in performance its classic analogue, and by further improvements it can be compared with higher quality learning algorithms.

2. Materials and Methods

2.1. Area of interest and air quality data

Thessaloniki is the political and financial capital of the Central Macedonia region in Greece for the last 2500 years. A dense urban web with high accumulation of urban activities characterizes the city, where traffic congestions are common during rush hours. The industrial area located to its west contributes to the air pollutant emissions resulting from the more than 1,000,000 inhabitants, and the overall air quality levels can be quite problematic especially when it comes to particles and gaseous pollutants like NO_2. Air quality data used in this study originate from the monitoring network operated by the relevant authorities. We conducted computational experiments focusing on the mean daily concentration of particulate matter (PM10) in the central area of Thessaloniki, where the Egnatia station is used as a reference. The available dataset contained 2556 records of mean daily values, which are characterized by nine input variables (features). Three of them are meteorological, i.e. mean daily temperature (°C), humidity (%) and wind speed

[1] Environmental Informatics Research Group, School of Mechanical Engineering, Aristotle University, Thessaloniki, Greece, arkaragi@auth.gr
[2] Environmental Informatics Research Group, School of Mechanical Engineering, Aristotle University, Thessaloniki, Greece, kkara@auth.gr

(m/s). The rest measure the mean daily concentration (μg/m^3) of various pollutants in Egnatia station and two other stations in the Thessaloniki area. As a target variable for our modelling efforts, we set the mean daily concentration of PM10 for the next day. The records represent days from 01/01/2007 to 31/12/2013. The first six years (2007-2012) defined the training set, while the year 2013 was used as the test set. All years were selected based on data availability and in an effort to minimize missing or problematic data.

2.2. Fuzzy decision tree

The algorithm that we used for the induction of the fuzzy decision tree is based on the work of Yuan and Shaw [3]. We propose a modification of the information measure used in the partitioning process by introducing a consciousness function [4] which merges the nonlinearity decision preferences into the decision making, thus improving the accuracy of the obtained knowledge. This is due to the nonlinear influences of the membership state in many decision-making problems. With given significant level α and truth level threshold β, the only hyper-parameters of the model, the induction process consists of the following steps:

Step 1: Fuzzify the numerical variables of the training set to linguistic variables. This process is carried out by employing the Fuzzy C Means clustering algorithm to define some prototypical trapezoidal forms based on given data and then make corrections to the parameters, in order to find those which minimize the error during the testing phase. Figure 1 presents the most significant linguistic variables.

Step 2: Measure the modified classification ambiguity associated with each attribute and select the attribute with the smallest classification ambiguity as the root decision node. The modified classification ambiguity is calculated by replacing membership functions with a function of the form

$$f(x;n) = \mu^n(x)$$

Step 3: Delete all empty branches of the decision node. For each nonempty branch of the decision node, calculate the truth level of classifying all objects within the branch into each class. The objects within each branch are the examples of the decision node with a membership degree greater than the significant level α to the fuzzy set that corresponds to the branch. If the truth level of classifying into one class is above a given threshold β, terminate the branch as a leaf. Otherwise, investigate if an additional attribute will further partition the branch (i.e. generate more than one nonempty branch) and further reduce the classification ambiguity. If yes, select the attribute with smallest classification ambiguity as a new decision node from the branch. If not, terminate this branch as a leaf. At the leaf, all objects will be labeled to one class with the highest truth level.

Step 4: Repeat step 3 for all newly generated decision nodes until no further growth is possible; the decision tree is then complete.

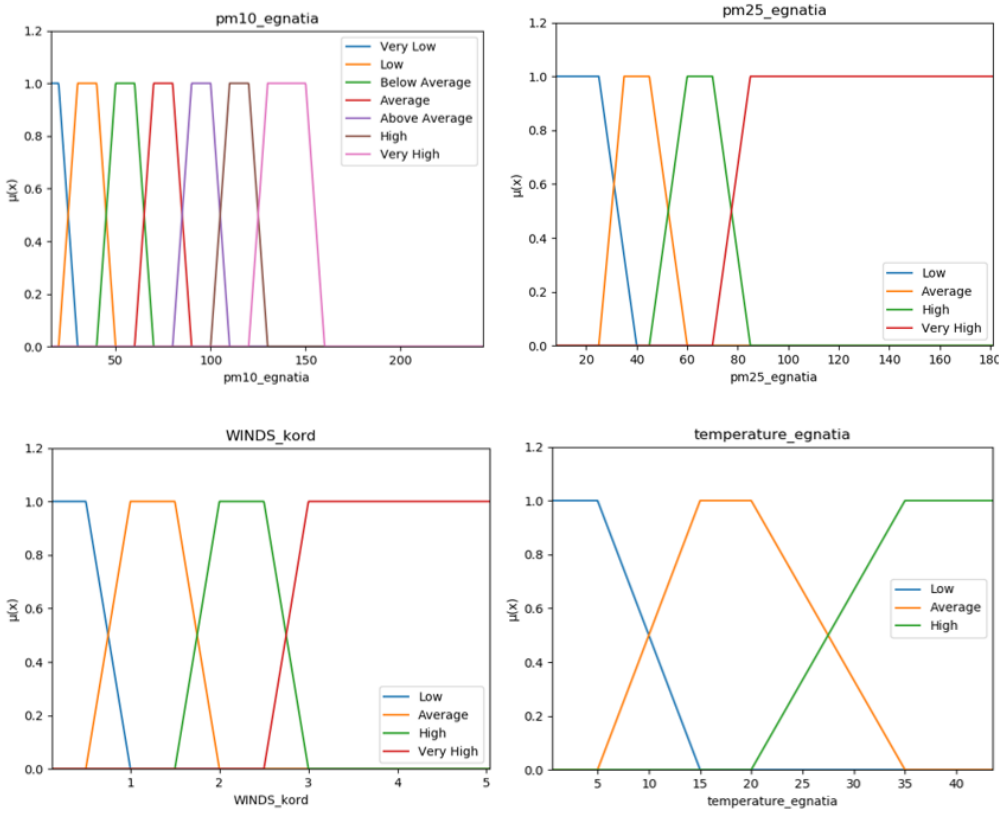


Fig. 1. Fuzzification of the numerical environmental variables to linguistic variables. The values of a linguistic variable are expressions of the physical language and are represented by fuzzy sets.

3. Results

The training set was used for the induction of a fuzzy decision tree and for the development of a data-driven model, with which it will be compared. This model is a random forest consisting of 300 decision trees. The hyper-parameters of the fuzzy decision tree were tuned after trial and error and were assigned the values α=0.15, β=0.5. At Table 1 we present the results of the computational experiments per model.

	R^2	RMSE	MAE
Fuzzy decision tree	0.444	0.080	0.056
Random forest	0.347	0.087	0.059

Table 1. Model performance statistics. R^2: coefficient of determination; RMSE: Root Mean Square Error; MAE: Mean Absolute Error.

Figure 2 presents the time series of the actual and the predicted values during the test phase. Figure 3 shows the goodness of fit of the fuzzy model. The poor performance of the model at some value regions is owed to examples which are classified based on a single rule or a couple of rules, which consequently produce the same fuzzy set. The first 13 fuzzy rules are included in Table 2.

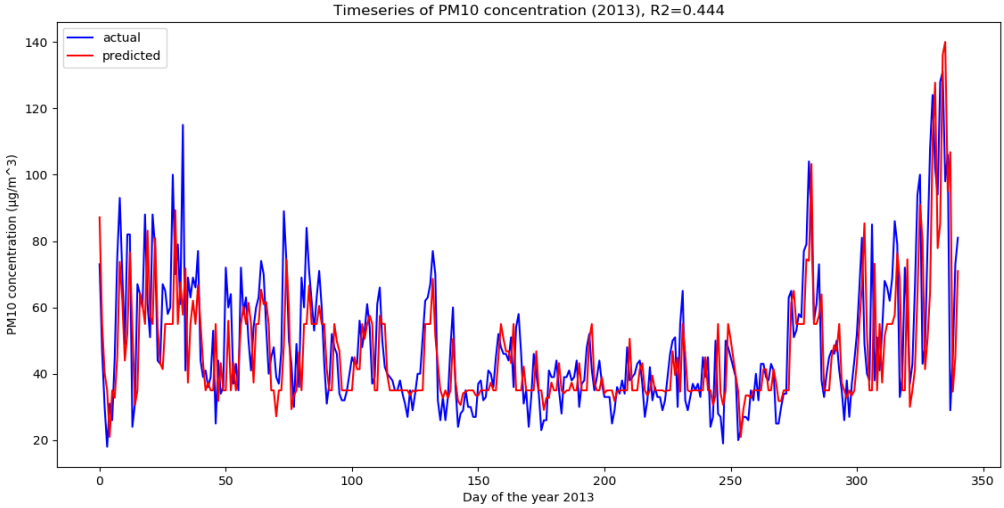


Fig. 2. Time series of actual values (blue line) and predicted values (red line).

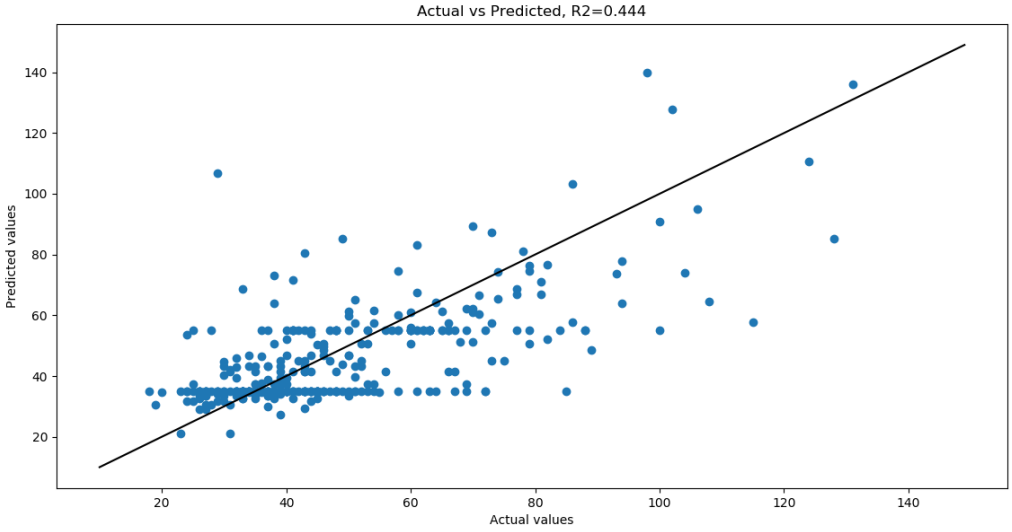


Fig. 3. Goodness of fit of the fuzzy model.

If pm10_egnatia is "Very Low" AND so2_egnatia is "Low" THEN Very Low (S=0.54)
If pm10_egnatia is "Very Low" AND so2_egnatia is "Average" THEN Low (S=0.53)
If pm10_egnatia is "Low" AND so2_egnatia is "Low" THEN Low (S=0.48)
If pm10_egnatia is "Low" AND so2_egnatia is "High" THEN Below Average (S=0.51)
If pm10_egnatia is "Below Average" AND WINDS is "Average" THEN Below Average (S=0.49)
If pm10_egnatia is "Below Average" AND WINDS is "Very High" THEN Very Low (S=0.53)
If pm10_egnatia is "Average" AND WINDS is "Very High" THEN Low (S=0.54)
If pm10_egnatia is "Above Average" AND WINDS is "Low" THEN Above Average (S=0.43)
If pm10_egnatia is "Above Average" AND WINDS is "Very High" THEN Low (S=0.63)
If pm10_egnatia is "High" AND WINDS is "Low" THEN High (S=0.47)
If pm10_egnatia is "High" AND WINDS is "Very High" THEN Low (S=0.61)
If pm10_egnatia is "Very High" AND WINDS is "Low" THEN Very High (S=0.54)
If pm10_egnatia is "Very High" AND WINDS is "Average" THEN High (S=0.59)

Table 2. Fuzzy rules extracted from the fuzzy decision tree.

4. Discussion

The results of the computational experiments prove that the fuzzy decision tree is an improved version of its classic analogue, outperforming a random forest in precision and offering many possibilities for further improvements. Fuzzy rules allow the extraction of reasonable conclusions from human beings, without the need of expertise in the field of application, since the concepts of Low and High are much closer to the human intuition, rather than a strict inequality, like these which express the rules of a classical decision tree. Fuzziness is responsible for smoothing the steep transitions between the subsets that occur when the input space is partitioned, allowing the gradual membership of every example to one or more subsets. Another advantage is that it can be applied to dataset without the need of clearing missing values and normalizing numerical variables. We notice that even with missing values in the training set the fuzzy tree shows improved efficiency over the Random Forest algorithm. On the other hand, there have been research results reporting a coefficient of determination higher than 0.8 for a number of machine learning algorithms when it comes to the mean daily PM10 concentration forecasting [5], thus suggesting that our research can benefit from a more thorough feature selection phase and a refined algorithmic testing phase.

The specific fuzzy decision tree algorithm is still at an early stage of development. Its performance depends on a multitude of parameters, which appear during the induction of the tree, the formation of fuzzy rules, and the decision-making process (defuzzification). Given the fact that the choice of the most suitable parameters for the optimization of the fuzzy model must depend on the nature of the problem being studied,

the development of an algorithm for the automatic adaptation of these parameters is the next research step. Some of the most significant parameters are: the choice of the fuzzy union and the fuzzy intersection operators, the choice of a proper index that measures the degree of subsethood, the replacement or the modification of the information measure for more informative partitions of the input space and the usage of a feedback method for calculating the optimal membership functions. We also propose the combination of fuzzy decision trees for implementing a fuzzy random forest.

References

[1] Quinlan, J. (1986): Induction of decision trees. In: Machine learning 1, 81-106.

[2] Zadeh, L. (1973): Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. In: IEEE Transactions on Systems, Man, and Cybernetics 3(1), 28-44.

[3] Yuan, Y.; Shaw, M. (1995): Induction of fuzzy decision trees. In: Fuzzy sets and systems 69, 125-139.

[4] Jin, C.; Li, F.; Li, Y. (2014): A generalized fuzzy ID3 algorithm using generalized information entropy. In: Knowledge-based systems 64, 13-21.

[5] Rybarczyk, Y.; Zalakeviciute, R. (2018): Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. In: Applied Sciences 8, 2570.

PigFarm
Developing Decision Support for the Pork Production Industry

Thomas Rose[1&2], Julia Gruber[1], Kathrin Gunkelmann[1]

1. Introduction

This work in progress reports on PIGFARM (Fact and Rule Management for improving decision support in pork production) as an integration framework for fusing information in the context of raising pigs and deducing rules to improve breeding procedures. Information about pork production is typically spread across different stakeholders such as farmers, veterinary, and slaughterhouses. Integration and feedback across these information islands is sparse. At best, a farmer might receive feedback form the slaughterhouses based on the inspections after slaughter. Such inspections deliver medical feedback, but feedbck limited to defined examination indices and no rationale for potentially bad performances are given. Well-being of animals during their life time can be concluded in principle, but conclusions for counter measures are sparse. However, a mathematical model of production parameters and animal well-being is an open challenge. Current models unveil impacts by qualitative correlations. Since no direct links to the production parameters are available in order to reflect a production information landscape as digital twin of a pig's life-cycle, causal relationships between well-being and the animal breed cannot be derived. Hence, quantitative evidence on the impact of production parameters such as keeping conditions on an animal's well-being is missing.

PIGFARM overcomes this limitation with a controlled knowledge cycle. Based on a fused set of data from several sources such as farm management, veterinary supervision and inspection at the slaughterhouse, rules for the impact of environmental conditions and production parameters on animal well-being will be derived. Such rules, i.e. impacts of decisions of production parameters on the well-being of the animal, allow farmers and farming consultants to predict the impact of changes in production parameters, e.g. change of stable floors or air condition. Slaughterhouses also put economic value on pigs delivered. This assessment is based on certain properties of the pork, such as distribution of fat versus ham, and the spread of weight for the entire group of pigs delivered. Hence, the farmer also receives feedback on yields achieved finally resulting in economic impacts of decisions drawn during the life-cycle. Since these rules provide valuable input for decision support, they will be traded by PIGFARM as goods and intellectual property respectively. The question arises, how to monitor the use of such rules and reimburse rule creators, i.e. sources of information supplies as well as analytic experts. Then, experiences gained with certain rules can be returned back to the information sources. Hence, our knowledge cycle resembles a control circuit. Pivotal element is the deduction of rules.

[1] Fraunhofer FIT, Schloss Birlinghoven, 3757 Sankt Augustin, Germany, thomas.rose@fit.fraunhofer.de
[2] RWTH Aachen University, Informatik 5, Ahornstraße 55, Aachen, Germany,
(julia.gruber|kathrin.gunkelmann)@fit.fraunhofer.de)

Our approach spans upon five pillars:

- A *data space* for the fusion of information streams originating from different stakeholders, while maintaining sovereignty of stakeholders is of pivotal importance. Integration and linking of sources are based on classification concepts of livestock trade and farm management such as VVVO number (Viehverkehrsverordnung – a German term for the identification of farms).
- A *statistical analysis* for the deduction of causal relationships between significant attributes in our space of instances and attributes. Such causal relationships define the rules for changing housing conditions and keeping parameters. They furnish the tools for farmers and farming consultants to improve pork production with regards to animal well-being and raising efficiency.
- *Portal* for publishing such rules in order to disseminate the knowledge of the statistical analysis. The portal merely displays the availability of such rules with some economic key parameters and mining evidence for quality assurance. Specific contents will only be available once registered.
- A *rule monitor* controls the deployment of the rules as well as the application of rules on particular farms and livestock each single time, that is, once purchased each rule use, i.e. knowledge application, will be honored towards the creators as incentive system.
- An *eco system* that furnishes the soil for the different stakeholders to cooperate. Different stakeholders must make information sources available and knowledge beneficiaries must return a share of their yields. Hence, an appropriate business model is instrumental for a successful collaboration of the different stakeholders.

Key element of our approach is a mathematical model for the causal relationships between keeping conditions and animal' health as determined by the slaughterhouse. That is, how to capitalize on the digital twin of production processes in the pork industry. The correlation between the health status of a livestock and animal welfare is unquestioned although the borderline has to be sharpened in an animal physiology perspective. Besides other factors, the reduction of diseases has proven to be instrumental for improving animal well-fare [1]. Qualitative checklist and audit reports with various disease indices provide one solution to assess well-fare. Analysis of medical inspections at slaughterhouses have been proposed as alternative. Our approach favours a quantitative stance.

This paper is organized as follows. Section 2 reflects the argument that any measure to improve animal well-fare also has a significant positive impact on the production process, i.e. the ratio of food invested and yields gained. There is no conflict of objectives between animals' well-being and economic efficiency of the breed. Then, our formal model is summarized in section 3 as anchor point for production governance while section 4 pencils the functional scope of PIGFARM, its technology ingredients, and achievements accomplished. Further challenges to be addressed in order to strengthen the entire eco-system are presented subsequently. Finally, we discuss the impact of digitalization of the pork industry on sustainability, i.e. how can the digital twin of the production environment as represented by our models improve farming to become not only smart but also ecologically sound.

2. Animal Well-being and Economic Efficiency – Two Twin Sisters

Major sources for the analysis of animal well-being are the veterinary inspection data from slaughter houses. They reflect objective assessments of animals' health status, although variations in assessment quality are common. However, these data are merely maintained on a group level, i.e. the pigs examined per delivery by farmers and varies noticeable due to diverging perceptions of veterinaries. Even the time of the day and the day of the week have an impact on the inspection results. Yet, some variations can be smoothened by normalization. There are no individual links between pigs and inspection results. All results are grouped according to delivery units.

Yet, there is strong evidence about the impact of diseases on production efficiency. A research project by IQ-Agrar unveiled such a relationship impressively [4]. Pigs have been monitored on an individual basis thanks to the use of RFID technology. Hence, they were able to relate production performance with diseases on an individual basis.

Production performance is typically metered in weight increases per day per pig while the typical ratio of investment and yield is 2 to 1, i.e. 100 grams of food offered translate to 50 grams weight increase gained. A study with 100.000 pigs in 2017 unveiled impressive correlations [4]:

- 74 grams loss of increase for pigs with a disease diagnosed;
- 100 grams loss of increase for pigs with more than one disease;
- 140 grams loss of increase for pigs with liver-related diseases and pleurisy, which is a specific inflammation near to the lung.

Test animals with diseases faced substantial stress due to their diseases suffering from pain and physical harm. The numbers unveil a tremendous loss of invested food compared to healthy pigs. Hence, health is not only good animal welfare but also a key performance indicator for production efficiency. Loss of production performance is basically due to the fact, that pigs with diseases require additional nutrition to stabilize their immune system. Moreover, in different life phases they build muscle meat and grease at different intensities. Losses worsen once caught in the wrong phase [6].

To re-iterate, health is certainly a desirable objective. Farming indices for assessing animal comfort on the farm have also evolved as a kind of marketing tool. Essentially, the question arises of how to predict potential threads of animals' health as major indicator for their health and welfare. A formal model of causal relationships between environmental conditions and health status is required. Besides matters of well-being, such model has certainly overwhelming repercussions, i.e. there is no conflict of objectives between health as key for welfare and economic yield. However, a quantitative model about causal relationships is needed rather than arguing with qualitative arguments in an eminence stance without veterinary evidence.

3. Formal Model of Animal Well-being for Production Control

The anchor point of our approach is a formal model for causal relationships between environmental conditions and health impact. As of now, farmers merely receive charts as feedback from the slaughterhouses as depicted in Figure 1. These charts taken from a portal of IQ Agrar[3] basically summarize major facts such as weight, yields gained due to the quality of the pork and the like. They rank farmer's performance with their colleagues. However, such rankings do not tell the farmer what to change in order to improve performance. A model with the causalities is missing.

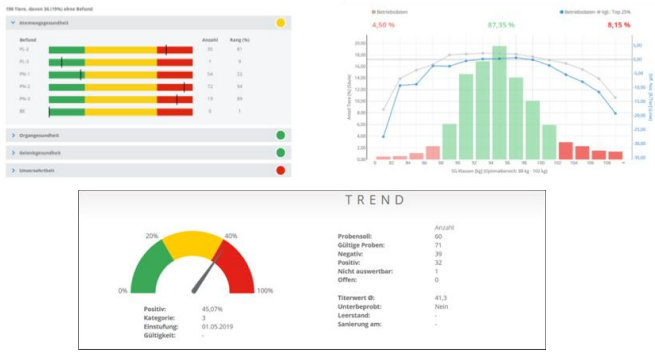


Fig. 1. Production performance in terms of rankings for key indicators.

That is why we started with project PIGTALE[4] to analyze potential impacts of environmental conditions such as stable design or air condition and animals' health status as inspected by veterinary officials in slaughterhouses. The medical inspection in the slaughterhouse is a public responsibility and governed accordingly.

We run different sprints to generate a model of causalities:

- integrate data sets from different sources and repair missing or incomplete attributes;
- depict the role of different attributes and identify significant ones for the mathematical model;
- initiate the analysis by exploring potential causalities.

Starting point for building such a statistical model is the data base. For combining information sources from various stakeholders like farmers, veterinaries or slaughterhouse the integration of all data sets is essential. Within the integration process different steps need to be passed through. A first step is the linking of common information of all data sets such as the VVVO number to generate one joint data set. In another step the intricate preparation process needs to be done, such as checking for missing data and whether those missing aspects have a specific meaning, compiling location factors and merging selected outcomes to particular clusters for in form and content reasonable variables.

After generating a useful joint data set, the research of possible connections between

- environmental conditions of the life cycle of a pig,

[3] https://www.iq-agrar.de/
[4] https://www.fit.fraunhofer.de/de/fb/risk/projects/pigtale.html

- the managerial aspects of the farmers business and
- the health conditions of the animals as determined in the slaughterhouses starts.

Based on domain competence of our collaborating farmers and veterinaries as well as knowledge from literature, correlations are identified and verified based on mathematical approaches like the forward selection, which also sets the variables in a decreasing order determining the impact on the dependent variable of interest [8] – in numerous cases a health aspect of the animal.

Knowing the connections between the variables of interest, building a statistical model is the final step in this process. Due to the complicated and time-dependent structure of the data, PIGTALE uses a non-linear Tobit regression model in order to derive causalities since the dependent variable ranges in a specific co-domain. We modeled the occurrence of lung diseases within one purveyance of animals in a binary way and calculated the frequency for each farmer.

With different robustness checks, such as fixed effects, which excludes non-observable time-constant characteristics [3], the last sprint for generating a model for causalities ends and we come up with a formal pattern to predict impacts. As shown in Figure 2 we are able to converge the diagnosed diseases with our predicted ones from the model.

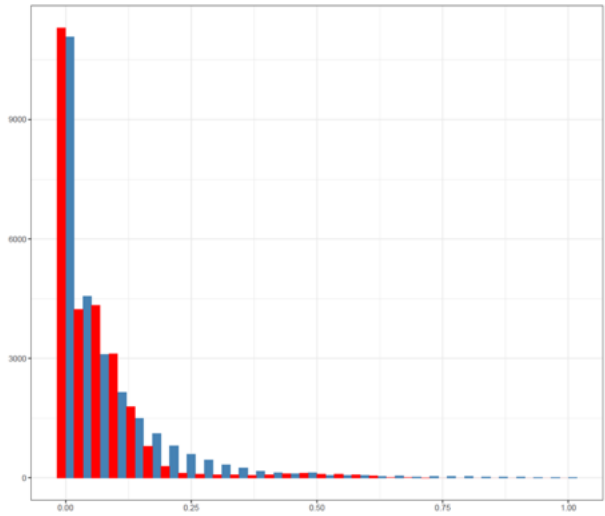


Fig. 2. Lung diseases predicted (red) versus diagnosed (blue) at slaughterhouse.

This statistical analysis has generated a formal model for causalities between environmental conditions and the health status of a pig. Given a set of living conditions and some parameters of the breed such as line of race, probabilities of diseases can be predicted. However, Figure 2 presents the formal model for lung diseases, which are certainly important but not the only disease of interest. Other diseases have to be analyzed accordingly.

Such rules for impact prediction support farmers in their decision processes, e.g., for installing an automated air condition or changing the floor of stable. Economic savings achievable due to health

improvements can directly be related to the costs of investment. Hence, investment decisions are evidence-based and no more blurred by only qualitative factors.

Since the statistical analysis shows limitations of environmental conditions such as housing or other process parameters, these insights can also be considered to improve audit reports in the farming industry. Such audits are typically been used for quality control and the identification of counter measures in case of open issues. In such cases, a statistical analysis can crosscheck the audit process to identify anomalies.

4. Functional Components of PIGFARM

PIGFARM comprises five major elements as already indicated (confirm Figure 3). The eco system is the organizational heart of our approach. It provides the soil for business collaboration among our partners stemming from different sectors of industry. The current version of our eco system is founded in a community model. Farmers, veterinaries, slaughterhouse and respective representatives teamed-up. They provide data and their expertise on a cooperative basis and are allowed to use the knowledge derived. Hence, partners operate as a kind of cooperative association. A Proof of Work has been provided successfully with our prototype and the development of business models is ongoing.

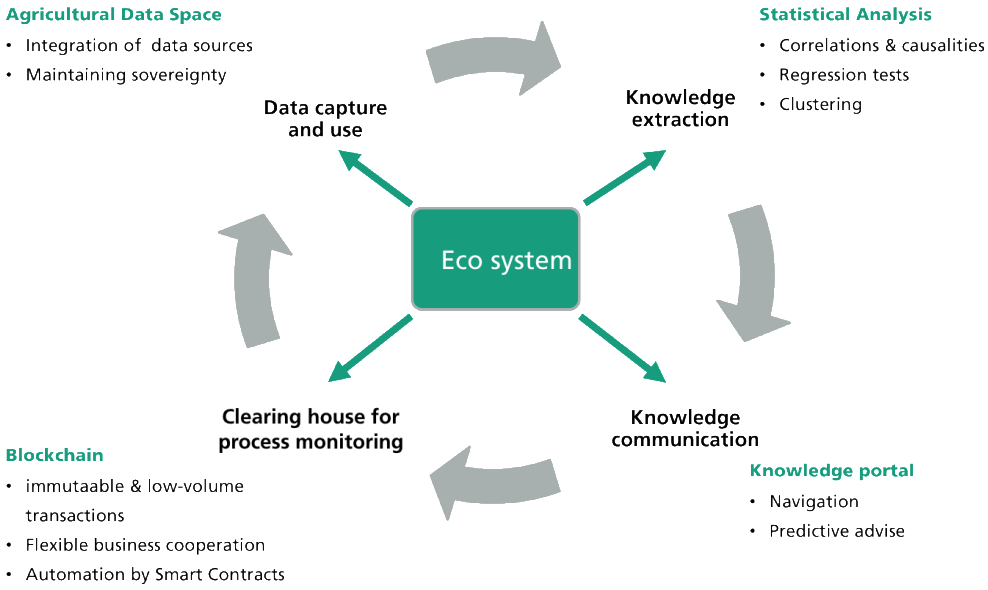


Fig. 3. Functional components of PIGFARM.

The anchor point of PIGFARM is the statistical analysis with its methods, tools and technologies. This element generates the intellectual capital for the pork production industry. This capital is traded via a portal and its use is monitored by the clearing house. Hence, use of this capital is documented by DLT [9] as immutable technology for revision safety.

Communication of knowledge derived is performed by the portal for knowledge transfer. The portal is not limited to reactive navigation support, but also offers predictive services to certain farmers, e.g., if the number of lung diseases is higher than usual, then counter measures can be offered.

The clearing house monitors the use of the rules for improving pork production processes. On the one hand, it monitors use of the rules. Each use is charged on a single-basis. Hence, billing is based on a use-oriented basis rather getting usage-rights on a subscriber or flat basis. On the other hand, it transfers incentives for using rules towards data providers and analytic experts. Hence, it settles demands between knowledge users and enablers for knowledge generation.

Thus, a controlled knowledge circuit is in place. It gathers data about production processes, analyses rationales for process performance and offers a framework for knowledge exploitation and rewards. The system is successfully proven by PIGTALE as already indicated. Open challenges revolve around

- data governance and its technical automation in an open and fair environment,
- business model for fair and sustainable operations, and
- a reward system to honor the quality of data provided and its contribution towards the quality of rules derived.

5. Open Challenges

Two major challenges are currently ahead for the governance of our eco system: consolidation of a data management framework and an incentive system as basis for a business model. Coordination of data management currently follows an organizational approach with procedures agreed for date provision, access and use. Hence, partners provide their data on the basis of certain legal agreements. Data access and distribution is governed by such agreements among organizations. A future extension for data capture is the deployment of the Agricultural Data Space (ADS) as instantiation of the Industrial Data Space [7]. The ADS will translate organizational procedures into technical services in terms of connectors. Such connectors govern access to data according to policies agreed. Hence, regulations will be translated into automated procedures to protect intellectual properties on data as well as sovereignty of data providers. Even use of data can be monitored, i.e. data used rather often in the analysis might receive higher incentives than others.

An incentive system has to address two dimensions: data capture and analytic services. Both lines of partners bring value to the eco system and thus deserve incentives. Any business model has to consider these dimensions. Yet, there is still an open issue: what is the specific value of the rules we generate. Some have a strong economic impact while others are rather intangible in particular when animal welfare is concerned. In addition, models for subscription services have to elaborated to address different scopes of use: how about temporary users of knowledge, users of knowledge only derived from a subset or users only interested in the impact of food from specific mills. Different kinds of data cubes in terms of attributes or instances might be used during the analysis perhaps resulting in varying qualities of rules derived. Figure 4 illustrates different modes of subscription and vibrant data scopes for the statistical analysis.

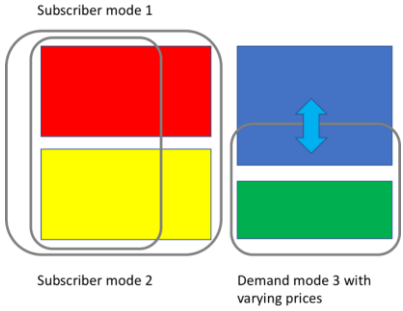


Fig. 4. Subscriber-oriented views and vibrant data cubes as sources for analysis.

Currently, all data provisions are accounted for on an equal basis. However, data sets might have different impacts on the quality of the analysis. Frequencies of use might differ as well as impact created by rule application. The clearing house can easily document them, but rewards are an open issue. Hence, the ADS and the clearing house ought to be used to monitor the use of data for analytic purposes. Interestingly, individual contributions of single data sets can also be different. Imagine a classifier in terms of a discriminator partitioning entities in good and bad. There might be a line to tell apart. Data far away from the border line might be less valuable than those very close to it, since they can sharpen the classification process.

Manual tools for data capture prevail. Sensor technology is certainly an interesting aspect to lower costs for data capture. Another aspect of data capture is consistency. The inspections of veterinaries at slaughterhouses are currently rather subjective than repeatable on an objective basis. The question arises of how to achieve harmonization. Optical recognition might be one solution towards normalization of inspections. Once captured by a camera – not exactly an easy thing in the environment of a slaughterhouse – an optical classifier can zoom to certain spots and organs. Then, it can be compared to common references and conclusions can be generated such as level of inflammation by 10%.

6. Conclusion

This progress report presents a digital twin of pork production processes spanning across the sovereignties of different stakeholders. The digital twin furnishes a platform to derive knowledge for improving the performance of production processes. One highlight is quantitative evidence in terms of a formal model founded in statistics to predict the impact of environmental conditions – such as housing condition, air conditioning or food deployed – on animal health and finally economic efficiency of the production process. Such a formal model gives all parties proven advise to optimize breeding conditions rather than relaying on qualitative knowledge.

Secondly, we developed a framework for capturing, trading, monitoring and rewarding this knowledge. Hence, we invented a new kind of good in pork production processes: the efficiency of change in keeping conditions. A farmer can now predict the changes conducted and thus gained a new level of control in managing production processes.

This kind of knowledge is certainly not limited to improving production processes, but can also be applied to other scenarios such as risc-oriented inspection at slaughterhouses. Veterinaries at the line in the slaughterhouse can look more carefully at spots with expected diseases or some inspections might be skipped because they can be excluded from the history of an animal's life cycle.

Acknowledgements

This work has been supported in part by the b-it foundation[5]. We would like also to thank our colleagues from our collaboration partners IQ Agrar, Osnabrück, Erzeugerring Westfalen, Senden, and QuH Lab, Siegen. Moreover, we would like to thank our teammates Marlene Bubb, Timo Meiendresch, and Thomas Osterland.

References

[1] Baltagi, B., Song, S., & Koh, W. (2003). Testing panel data regression models with spatial error correlation. Journal of econometrics, 1, S. 123-150.

[2] Blaha, T., & Sundrum, A. (2017). Tierärztliche Kompetenz und Zielorientierung erforderlich! Deutsches Tierärzteblatt, 65(11), S. 1518-1521.

[3] Greene, W. (2004). Fixed effects and bias due to the incidental parameters problem in the Tobit model. Econometric reviews, 23.2, S. 125-147.

[4] Hartmann, F. (2018). Jahresbericht IQ-Agrar 2017.

[5] Henningsen, A. (2010). Estimating censored regression models in R using the censReg Package. (R. p. vignettes, Redakteur)

[6] Klauke, T., Piñeiro, M., Schulze-Geisthövel, S., Plattes, S., Selhorst, T., & Petersen, B. (2013). Coherence of animal health, welfare and carcass quality in pork production chains. Meat Science, 95(3), S. 704-711.

[7] Otto, B., Auer, S., Cirullies, J., Jürjens, J., Menz, N., Schon, J., & Wenzel, S. (2016). Industrial Data Space-Digitale Souveränität über Daten. (F.-G. z. eV, Hrsg.) München.

[8] Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. Econometrica, 26, p. 24-36.

[9] Prinz, W.; Rose, t.; Osterland, T.; Putschli, C. (2017): Blockchain – Reliable Transactions. In R. Neugebauer, Hrsg.: Digital Transformation, Springer Vieweg, 301-309.

[5] http://www.b-it-center.de

Automated invasive alien species recognition: lessons learned from applying the iNaturalist 2017 computer vision model on citizen-science data

Blagoj Delipetrev[1], Sven Schade[2], Irena Mitton[3], Fabiano-Antonio Spinelli[4]

1. Introduction

Our planet hosts many millions of plants and animals, from which many might look identical. Due to their visual similarity, an expert is needed to classify species in the natural world. In the last decade, Computer Vision (CV) has made significant advancement because of deep learning (DL) algorithms [1]. The image classification of ImageNet dataset precision has risen from 50% in 2011 up to around 88% for Top-1 and 98% of Top-5 accuracy. Top-1 accuracy is the conventional accuracy, which means that the model answer (the one with the highest probability) must be exactly the expected answer. Top-5 accuracy means that any of the 5 highest model probability answers must match the expected answer. The iNaturalist Competition is a large-scale classification competition where the goal is to improve the state of the art in automatic image classification for real-world data [2]. The dataset contains 5,089 species, with a combined training and validation set of 675,000 images that have been collected and verified by multiple users at inaturalist.org.

The objective of the work presented here is to explore the possibility to enhance an existing tool (a mobile application/app) to recognize Invasive Alien Species (IAS) in Europe with CV capabilities based on the iNaturalist 2017 CV model. The European IAS dataset contains 1192 images of 59 distinct IAS gathered from 696 submitted observations. The performed analysis showed that the CV is capable of recognizing 18 IAS. The sub-dataset of these 18 IAS contains 65 quality-controlled images. The iNaturalist 2017 CV model processed these 65 images and produced results with 35.4% Top-1 and 47.7% Top-5 accuracy. The IAS recognition results are visualized and explained. The paper discusses how to include the CV in the IAS in Europe App as defined in Section 2, explores CV model fine-tuning strategies, discusses the current dataset limitations, and suggests possible future cooperation with the wider scientific community. The applied Python code is freely available on GitHub [3].

2. Invasive alien species in Europe

Recognizing the growing threats of IAS to human health, the environment and the economy [4], as well as, the needs for high quality data, we are committed to advance the knowledge base for IAS early warning, monitoring and management. For the work of the European Commission's Joint Research Centre (JRC) this is particularly related to the European Union (EU) regulation no. 1143/2014 "Invasive alien species of

[1] European Commission – Joint Research Centre, Ispra, Italy, blagoj.delipetrev@ec.europa.eu
[2] European Commission – Joint Research Centre, Ispra, Italy, s.schade@ec.europa.eu
[3] European Commission – Joint Research Centre, Ispra, Italy, irena.mitton@ext.ec.europa.eu
[4] European Commission – Joint Research Centre, Ispra, Italy, fabiano-antonio.spinelli@ext.ec.europa.eu

European Union concern" and the European Alien Invasive Species Network (EASIN[5]) - the official information system supporting the implementation of this EU Regulation. In order to increase information about IAS in Europe, we developed an approach to engage citizens in its collection and validation [5]. The current approach faces a particular bottleneck in species identification, which so far, remains a manual task that requires the knowledge of human experts.

Recognizing the increasingly serious problem of IAS in Europe, the European Commission published a dedicated Regulation [6]. The Regulation foresees three main types of interventions: prevention, early detection and rapid eradication, and management. It gives priority to a subset of IAS, included in the list of IAS of Union concern. Species are included in this list inter alia because they can cause such a significant damage in Member States justifying the adoption of dedicated measures at EU level. The implementation of this Regulation is supported by an official information system developed by the Joint Research Centre (JRC) [7,8], which collects scientific information and spatial data on IAS in Europe, supporting Member States competent authorities in the implementation of the Regulation on IAS.

The Invasive Aliens Species in Europe App originated from the MYGEOSS project [9], which has received funding from the European Union's Horizon 2020 research and innovation program, and was then taken up as part of the institutional work of the European Commission in house scientific and knowledge management service: the JRC. The project aims at developing smart Internet applications to inform and engage European citizens about the changes affecting their environment, and extend the pool of open source software and open data available to the global community through the Global Earth Observation System of Systems (GEOSS). The mobile app "Invasive Alien Species in Europe" is available for Android and iOS. The App is using available services to update the IAS catalogue, in that way the App is always up to date with the IAS catalogue of EU concern.

3. Methods

3.1. Invasive alien species dataset

The EU has currently identified 66 IAS of Union concern. The 66 IAS are divided into 30 animals and 36 plants. The 30 IAS of the animal kingdom are classified in 5 subcategories: 1 Amphibians (Amphibia), 5 Birds (Aves), 8 Invertebrates (Arthropoda), 11 Mammals (Mammalia), 1 Reptile (Reptila) and 4 Fishes (Actinopterygii).

Through the IAS in Europe mobile App there have been reported 696 observations consisting of 1192 images. In the 696 observations there are present 59 distinct IAS. 47 IAS are of Union Concern and 9 IAS of the local area of the Danube River Basin and 3 IAS from a testing area. The Danube River Basin local catalogue consists of 64 local IAS, none of them matching the iNaturalist 2017 list of species.

After the user submits observations of the species with the images (maximum of 3 images) the validation process takes place by a scientist who analyses the sent data and validates the observations. When an observation is submitted, the observation gets the status "Submitted" and in the validation process the scientist may change the status to:

- "Validated", if an observation with the images represent the reported species.

[5] https://easin.jrc.ec.europa.eu

- “Prevalidated”, if an observation image(s) represent the species but needs more checking.
- “Unclear”, if the observation image(s) is not clear and further information needs to be collected from the user.
- “Discarded”, if the observation image(s) are not representing an IAS selected.

On 9 March 2020, out of 696 observations, 399 observations were “Validated” as positive - the observation representing the declared species, 1 observation “Pre-validated”, 77 observations “Uncleared” and 217 observations “Discarded” as shown in Figure 1.

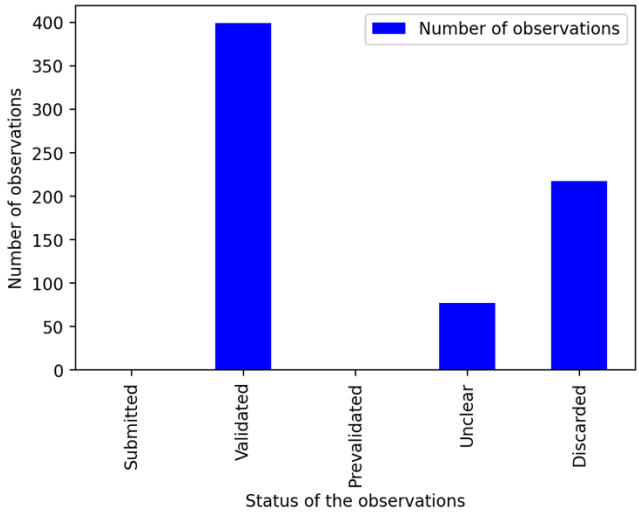


Fig. 1. Number of observations based on the validation status (9 March 2020).

The IAS dataset has in total 1192 images covering 59 distinct species. The IAS dataset has a long tail distribution, as shown in Figure 2.

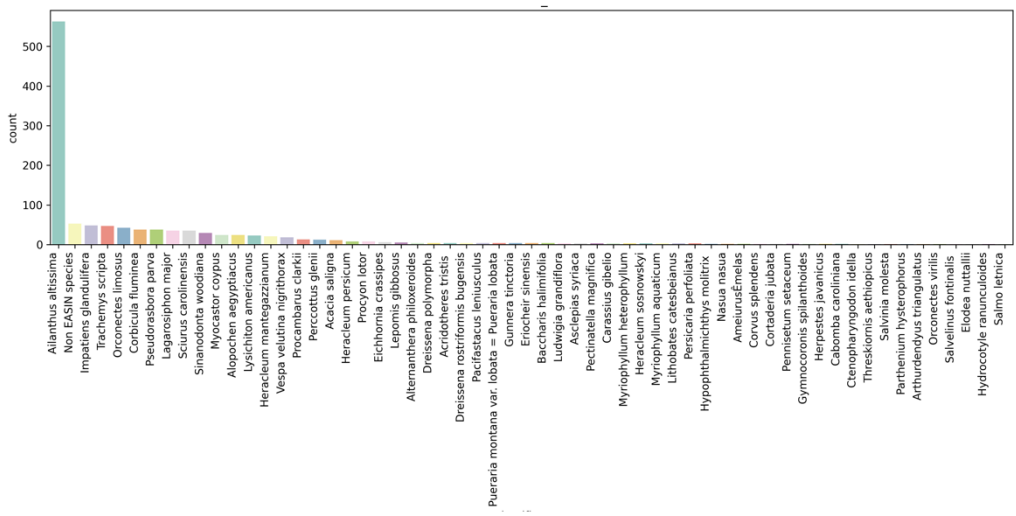


Fig. 2. Distribution of IAS images among species (9 March 2020).

3.2. Cross-checking of the iNaturalist 2017 and IAS datasets

The iNaturalist Challenge 2017 is a large-scale classification competition sponsored by Google and it was part of the FGVC4 workshop at CVPR 2017[6]. The goal of this competition is to push the state of the art in automatic image classification for real world data that features fine-grained categories, big class imbalances, and large numbers of classes. The dataset features many visually similar species, captured in a wide variety of situations, from all over the world. Example images are provided along with their unique GBIF ID numbers [10].

We have chosen the iNaturalist 2017 CV model because it covers a wide range of species and is one of the most popular CV models. There are following competitions in 2018 and 2019 have the same characteristic of long tail distribution where the majority of species have a small number of images, which is the same case with the IAS dataset.

To test CV capabilities in recognizing IAS, we made a cross-check analysis of the two datasets and investigated which IAS are supported by iNaturalist 2017 model. There are 18 IAS found with the cross-check in both datasets, as shown in Figure 3. The rest of the 41 IAS are not found and were therefore not included in the further analysis. The IAS dataset of 18 species has 148 images with different status e.g. submitted, validated, etc. (see also Figure 1) from which only 65 validated images of 9 IAS were used to test CV model.

We have developed an API on top of a Flask web service with PyTorch iNaturalist 2017 CV model [11]. The API receives the IAS images and sends them to the Pytorch model that returns five predicted

[6] https://sites.google.com/view/fgvc4/home

species labels with their probabilities. A script was created to automatically send the 65 IAS images to the API and store the results in a results dataset.

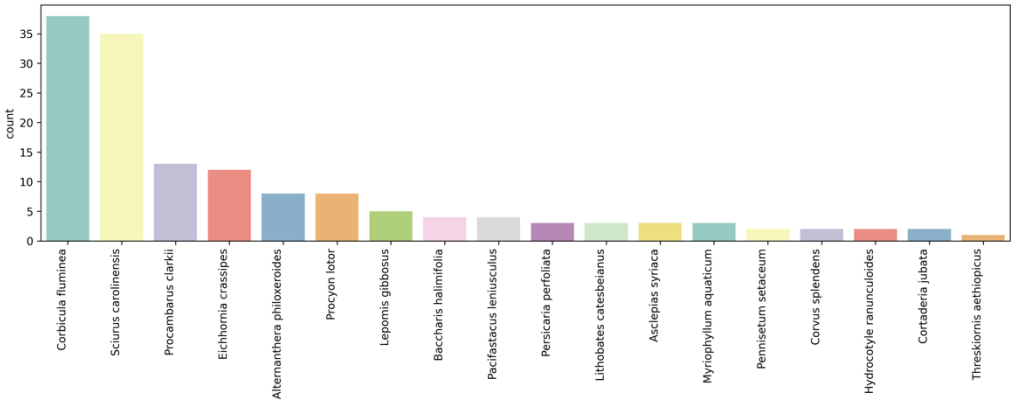


Fig. 3. Cross-checked species between IAS of Union concern and iNaturalist 2017.

3.3. Results of applying iNaturalist 2017 computer vision model to IAS dataset

The iNaturalist 2017 CV was successful in recognizing 35.4% Top-1, (35.4%+12.3%) 47.7% Top-5 and 52.3% of the 65 IAS images were not recognized, as shown in Figure 4.

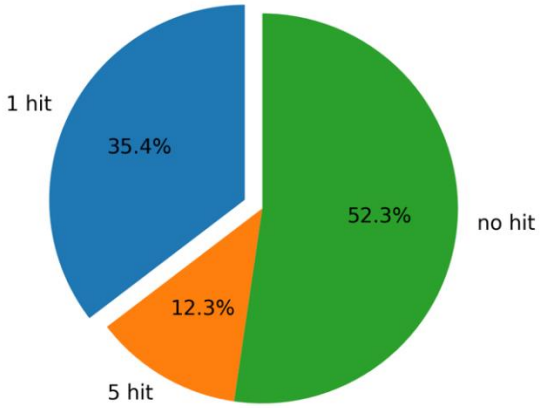


Fig. 4. iNaturalist 2017 CV recognition model on IAS dataset prediction results.

Examples of the iNaturalist 2017 CV model recognition of IAS images as shown in Figure 5. The first column includes IAS images, the second column are Top-1 predictions while from third to final column are the next 4 (in total Top-5) predictions. The probabilities are provided below each of the predicted images.

We developed a Python script to download example images from their unique GBIF ID numbers. However, many of the original links were broken and the sample image were not obtained and were marked with text "No image available".

In the first row of Figure 5, the IAS image is successfully recognized with a Top-1 prediction probability of 77.9%. The second row IAS is not recognized, and all predictions are below 10%. The third row IAS image is correctly recognized in the second prediction with 15% and it is counted in Top-5. The fourth row IAS image is not recognized and all prediction probabilities are low. The differences between the first prediction and the original image in the fourth row is very obvious. The fifth row is Top-1 prediction with probability of 100%.

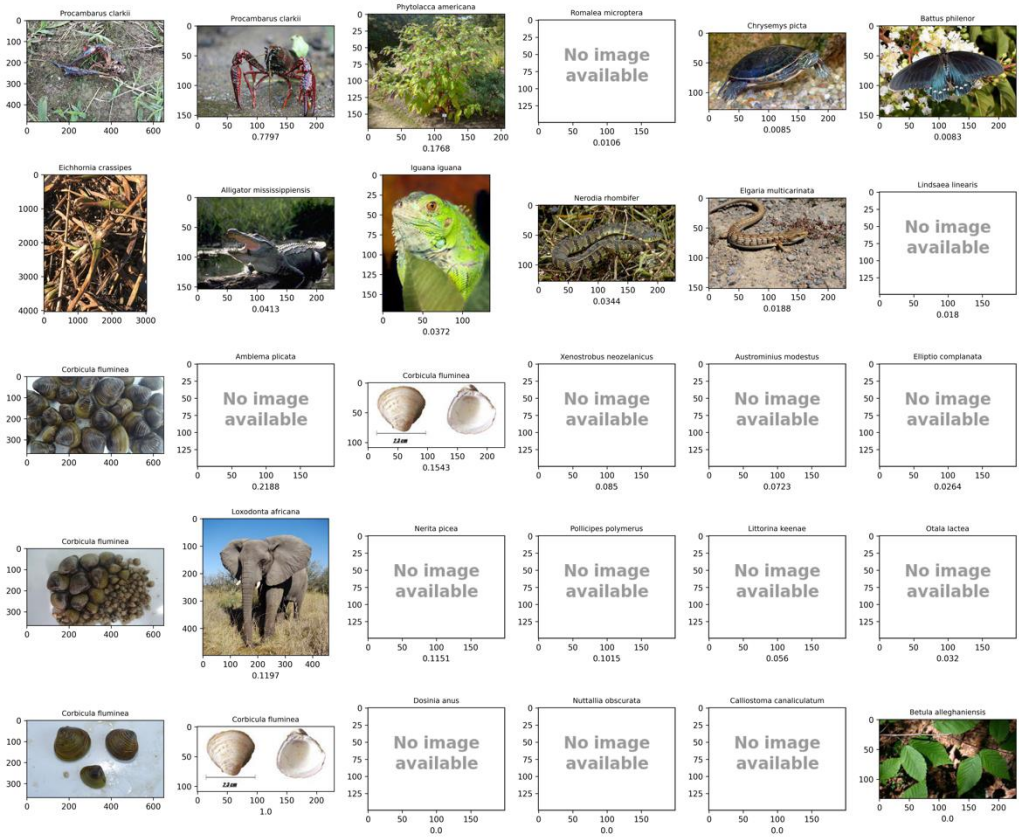


Fig. 5. Examples of iNaturalist 2017 CV model recognition of IAS images.

4. Discussion and lessons learned

Following our experiment, we found that developing an IAS CV model based on iNaturalist 2017 or any other CV model will depend on several factors:

- CV models apply a closed world assumption, i.e. they are trained for a particular set of species and will only be able to calculate similarities of a given image to the species that are parts of the model. If a species in an image is not available inside the model, the results will not help its identification.
- We need to further investigate how species identification depends on the image quality (esp. how the species features on the image) and the amount of images available for a single sighting. This investigation will help us develop user guidelines for taking pictures.
- We had only a small IAS dataset available. There is a need to establish cooperation between different initiatives and research groups, so that validation and test datasets can be created, and more advanced prediction models can be shared. A promising network to establish such collaborations is the COST Action AlienCSI [12]. Alternatively, augmented data could be produced based on the current IAS dataset.
- The iNaturalist 2017 model needs to be fine-tuned on the "enlarged" IAS dataset to improve model predictive capabilities. The fine-tuned model needs to include IAS that are currently not supported. It is also important that the IAS CV model does not lose its capability to identify a wide range of species.
- The newly developed IAS CV model will have to be integrated into the IAS in Europe app, so that we could directly benefit from these newly emerging capabilities.
- Last but not least, CV might be considered as a valuable part of validation procedures and quality assurance procedures. It should be combined with additional intelligence, e.g. derived from the space and time of the occurrence, known habitats of a species etc. We would consider it particularly valuable to include CV as part of the quality assurance process, but to embed it into a possible dialogue with the original observer of the species (used of the app in our case) or the entire community.

5. Conclusion and next steps

In order to improve the recognition capabilities of this model we are now looking into the development of guidelines for taking pictures of IAS, which also account for the difference between diverse species. Those guidelines will also include requests to provide more than a single picture for each sighting. Future tests will show if the guidelines prove useful, and if they will be taken up by all users.

Additional research is needed to develop a custom IAS CV model that will have better performance on IAS images while not losing the capabilities to identify a wide range of species as the ones included in iNaturalist. Here, we see the most value to establish international collaborations and jointly develop a reference database for training and testing models, as well as for model sharing.

Looking further into the future, we plan to exploit guided dialogues between the CV driven species recognition algorithms and the users of the App (or even the entire user community). On the one hand, we see such dialogues as an essential part of a collective intelligence approach to species identification. On the other hand, this method would also allow deliberations about the use of Artificial Intelligence (AI) and to that end would contribute to requests for people-centered and explainable AI.

Acknowledgments

The views expressed are purely those of the authors and may not in any circumstances be regarded as stating an official position of the European Commission.

The authors would like to thank Alexander Kotsev from European Commission, Joint Research Centre, for his helpful advice and comments that helped improve and clarify this paper.

References

[1] LeCun, Y.; Bengio, Y.; Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[2] iNaturalist 2017 competition (2017): https://www.kaggle.com/c/inaturalist-challenge-at-fgvc-2017 (20.07.2020)

[3] Research code GitHub (2020): https://github.com/deblagoj/iNaturalist-API (20.07.2020)

[4] Williamson, M. H.; Fitter, A. (1996). The characters of successful invaders. Biological conservation, 78(1-2), 163-170.

[5] Schade S.; Kotsev A.; Cardoso A.C; Tsiamis K.; Gervasini E.; Spinelli F.; Mitton I.; Sgnaolin R. (2019). Aliens in Europe. An open approach to involve more people in invasive species detection. Computers, Environment and Urban Systems, Volume 78, November 2019, 101384.

[6] Regulation (EU) No 1143/2014 of the European Parliament and of the Council of 22 October 2014 on the prevention and management of the introduction and spread of invasive alien species (2014): https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:JOL_2014_317_R_0003 (20.07.2020)

[7] EASIN - European Alien Species Information Network (2020): https://easin.jrc.ec.europa.eu/easin (20.07.2020)

[8] Invasive Alien Species Europe App: https://digitalearthlab.jrc.ec.europa.eu/sites/default/files/invasivealienspecieseurope_app_userguide.pdf (20.07.2020)

[9] MYGEOSS project (2015): http://digitalearthlab.jrc.ec.europa.eu/mygeoss (20.07.2020)

[10] iNaturalist Competition 2017 Competition (2017): https://github.com/visipedia/inat_comp/blob/master/2017/README.md (20.07.2020)

[11] iNaturalist 2017 Species Classification Challenge (2017): https://medium.com/@macaodha/inaturalist-2017-species-classification-challenge-4ff4d1499279 (20.07.2020)

[12] ALIEN CSI COST Action CA17122: https://alien-csi.eu/ (20.07.2020)

PART III
SENSORS AND INTERNET OF THINGS

$PM_{2.5}$ low-cost sensor performance in ambient conditions

C. Falzone[1], A.-C. Romain[1], S. Guichaux[2], V. Broun[2], D. Rüffer[3], G. Gérard[4], F. Lenartz[4]

1. Introduction

The use of low-cost air quality sensors to evaluate personal exposure, complement a monitoring network or perform real-time data assimilation is spreading. Although some laboratories and project consortiums have set up their own procedures to assess the performan ce of such systems [1, 2] and although the CEN TC264 WG42 is preparing two standards on that topic, so far the only European reference remains the Guidance for the Demonstration of Equivalence of Ambient Air Monitoring Methods [3] that holds for any kind of device, whether it costs 100 or 10 000 €.

This short paper presents the application on a recent data set of some generic statistical tests and the demonstration of equivalence to evaluate the performance of various air quality nodes. The demonstration of equivalence uses an orthogonal regression whereas a majority of papers present a linear model based on total least square [4, 5]. In Section 2, we present the measurement campaign set up, the devices and their main characteristics, as well as the metrics used to assess their performance. In Section 3, we show and discuss the results, whereas in Section 4, we draw our conclusion.

2. Material and methods

In this study, six devices based on low-cost sensors and designed by three different institutions are compared during three measurement campaigns held in January, February and April 2020. Their reliability is simply estimated in terms of data coverage, while their metrological performance is assessed on one hand by a comparison of statistical test results, error metrics and method agreement analysis, on the other hand by following, as closely as possible, the methodology for the demonstration of equivalence known and used by monitoring network managers.

2.1. Campaign

The measurement campaign set up for this study is an add-on to the one led annually by the Air Quality Department of ISSeP to verify and/or update the calibration factors of their PM_X monitors. In this work, three sites are investigated, the urban background station of Herstal from January 25th to February 6th, the suburban background station of Angleur from February 8th to February 20th and the temporary traffic station

[1] SAM – ULiège, Arlon (Belgium), cfalzone@uliege.be
[2] CECOTEPE, Seraing (Belgium)
[3] Sensirion AG, Stäfa (Switzerland)
[4] ISSeP, Liège (Belgium)

of Charleroi from April 10th to April 22nd. This last site is investigated during the Belgian lockdown due to COVID19.

2.2. Measurement systems

The sampler used for the reference gravimetric method is a conditioned Derenda PNS 16-6.1, while the instrument used as the equivalent method is the Grimm EDM180. The calibration equation in use since 2010 for $PM_{2.5}$ is $y_{cal} = \frac{y_{raw} - 4.256}{1.034}$. Nevertheless, for our tests we decide to consider the Grimm like the other devices and use the raw data. Gravimetric data are only available for the sites of Angleur and Charleroi, because a PM_{10} head was used in Herstal.

One commercial and five non-commercial systems based on low-cost commercial sensors are tested: the Nubo from Sensirion AG (3 replicates), the EcoCityTool v.2 (1) and v.3 (1) from ULiège and the Antilope v.3 (3) and 4 (3) as well as the Saiga (3) from ISSeP/CECOTEPE. Their main characteristics are reported in Table 1. All these automatic mini-stations are based either on the Honeywell HPMA115S0 or on the Sensirion SPS30 sensors for $PM_{2.5}$. The principle of measurement, light-scattering, is similar for both but they differ in the way they estimate mass concentration based on count concentration. The Grimm is also based on light-scattering but has a larger range and discriminate 32 particle sizes.

Model	Characteristics	Values
NUBO	Recorded parameters	$PM_{2.5}$ (SPS30), PM_1, T, RH, Td
(N032; N053, N111)	Communication	2G, 3G, 4G
	Developer	Sensirion AG
ECT v.2	Recorded parameters	CO, NO, NO_2, O_3, $PM_{2.5}$ (HPMA115S0), PM_{10}, T, RH
(ECT02)	Communication	2G
	Developer	ULiège – SAM
ECT v.3	Recorded parameters	CO, NO, NO_2, O_3, $PM_{2.5}$ (SPS30), PM_1, T, RH
(ECT03)	Communication	2G
	Developer	ULiège – SAM
Antilope v.3	Recorded parameters	NO, NO_2, O_3, $PM_{2.5}$ (HPMA115S0), T, RH, p, location
(An310, An38, An39)	Communication	Bluetooth
	Developer	ISSeP/CECOTEPE
Antilope v.4	Recorded parameters	$PM_{2.5}$ (SPS30), T, RH, p
(An411, An417, An47)	Communication	-
	Developer	ISSeP/CECOTEPE
Saïga	Recorded parameters	NO, NO_2, O_3, $PM_{2.5}$ (SPS30), T, RH, p
(Sa002, Sa003, Sa005)	Communication	2G
	Developer	ISSeP/CECOTEPE

Table 1. Automatic mini-station's main characteristics.

2.3. Performance evaluation

The comparison of medians is done through the non-parametric Kruskal-Wallis test, which requires neither normality nor homoscedasticity, and which can be applied to more than two samples. The comparison by pair is done through the non-parametric Mann-Whitney-Wilcoxon test with the Holm correction of the p-value. Accuracy is assessed via the computation of the Mean Absolute Error (MAE) and the Root-Mean-

Square Error (RMSE). Both error metrics provide a value in the same units as the original signals, the latter giving a higher weight to larger errors. Agreement is evaluated by computing Spearman's rank correlation coefficient and Lin's concordance correlation coefficient, and by making a Bland-Altman analysis. Finally, the verification of equivalence is done based on a limit value of 25 µg/m³, the relative expended uncertainty (W_{CM}) is set to 25% (uncertainty for fixed measurements according to the 2008/50/EC directive; it is set to 50% for indicative measurements) and the reference squared uncertainty equals 0.2221 $(\mu g/m^3)^2$. Depending on the results, a calibration will be required for some devices to reach the data quality objective. All data analyses are performed with R and the following libraries: openair, blandr, DescTools and Metrics.

3. Results

All low-cost measurements presented here have a 1-minute interval. When compared to the Grimm they are averaged with a 30-minute interval and when compared to the gravimetric method with a 1-day interval. For these aggregations a 75% capture rate is required, otherwise the value is considered as not available.

3.1. Preliminary tests

As a first step, we check data availability and discard instruments for which the coverage rate is less than 75% in each period (see Table 2).

	Herstal			**Angleur**			**Charleroi**				
	n	**%**	**[min-max]**	**n**	**%**	**[min-max]**	**n**	**%**	**[min-max]**	**n total**	**% total**
An310	19067	100	[0-93.5]	18735	100	[0.3-45.4]	18747	100	[3.2-58.4]	56549	100%
An38	19067	100	[0-97.7]	18735	100	[0.2-43.3]	18747	100	[2.7-61.8]	56549	100%
An39	19066	100	[0-194.6]	0	0	-	18747	100	[1.9-119.4]	37813	67%
An411	0	0	-	18699	100	[0-31.2]	0	0	-	18699	33%
An417	19030	100	[0-103]	0	0	-	18710	100	[0.8-68.4]	37740	67%
An47	0	0	-	18699	100	[0-29.4]	18710	100	[0.7-73.5]	37409	66%
Sa002	10747	56	[0.3-53.3]	13110	70	[0.7-31]	4702	25	[2-19.6]	28559	51%
Sa003	13730	72	[0.2-46.7]	10575	56	[0.6-24.3]	18165	97	[1-57.5]	42470	75%
Sa005	16785	88	[0-98.8]	17830	95	[0.5-21.2]	18174	97	[1.3-60.1]	52789	93%
N032	18801	99	[0.1-111.6]	18434	98	[0.2-29.9]	18535	99	[1.7-58.6]	55770	99%
N053	18758	98	[0.1-106.2]	18293	98	[0.2-25.8]	18408	98	[1.7-55.7]	55459	98%
N111	18764	98	[01-109.6]	18557	99	[0.3-32.5]	18594	99	[1.7-54.8]	55915	99%
ECT02	14525	76	[0-75]	18457	99	[0-26]	13620	73	[1-155]	46602	82%
ECT03	19067	100	[0-79.1]	18735	100	[0.2-46.4]	0	0	-	37802	67%

Table 2. Number of data, coverage percentage and minimal and maximal values collected during each campaign and for all of them.

As a second step, we make a visual inspection of the time series and a summary of the distribution via boxplot and whiskers.

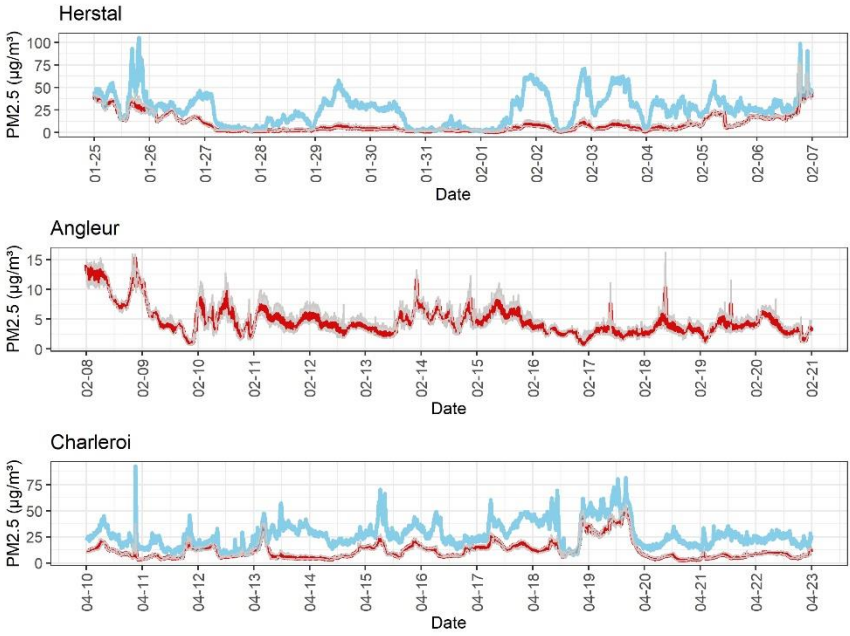


Fig. 1. Time series of the 14 instruments tested along the reference and equivalent methods. The red line is the median of the ensemble of instruments, the grey ribbon represents its interquartile range and the lightblue line the An39 device.

As can be seen in Figure 1, all sensors display a similar behavior over time for each period, except for the An39 (lightblue). However, the amplitude of the signals varies slightly from one sensor to the others as can be seen in Table 1. The levels observed in Herstal are below 20 µg/m^3 except during the first two and last two days, where they reach up to 50 µg/m^3. In Angleur, concentrations remain low (< 15 µg/m^3) during the whole campaign. For both these periods we have measured an amount of about 50 mm of precipitation, while the average wind speed was slightly higher for the second period than for the first, with respectively 4.79 m/s and 3.23 m/s. During the last period in Charleroi, a more diverse profile of concentrations is observed, e.g. a narrow peak on April 13th night, some days with an increase during the night and a decrease in the late morning, a whole day with $PM_{2.5}$ concentrations higher than 20 µg/m^3 on April 19th.

In Figure 2, one can clearly see the dependence of the distribution on the sensor model, at least for the first two boxplots. Both in Herstal and Angleur, the An310, An38 and ECT02, all equipped with a HPMA115S0, present a median higher than the one displayed by the devices with the SPS30 and also closer to their mean. The inter-quartile ranges are very similar in each device family.

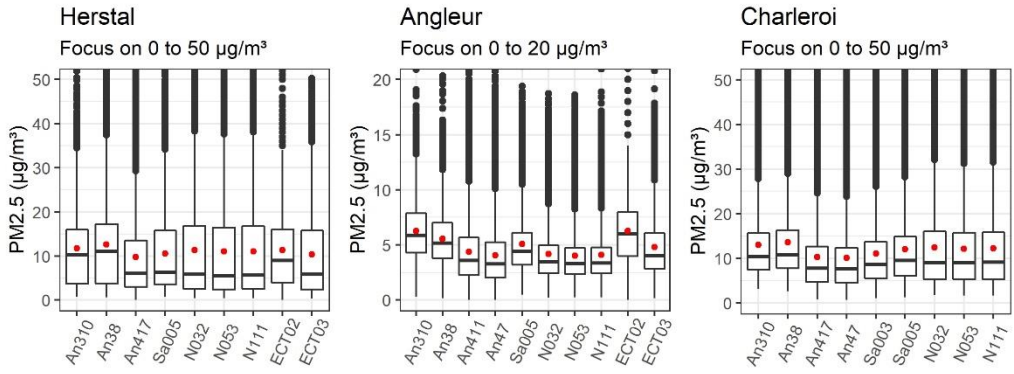


Fig. 2. Boxplots of the different devices for which the coverage rate is greater than 75% in each period and for a common period between devices. The red dot corresponds to the mean.

3.2. Statistical tests, error metrics and equivalence

The statistics tests are in accordance to the boxplots shown in Figure 2. The Kruskal-Wallis test applied on the ensemble of devices for each period presents a p-value < 0.05, hence the H_0 "Samples are from identical population" has to be rejected. The Mann-Whitney-Wilcoxon tests present a p-value ≥ 0.05 for some pairs, mostly those with the same SPS30 sensor (Herstal: An417 with the 3 Nubo and ECT03, ECT03 with the 3 Nubo and the 3 Nubo with each other; Angleur: An310 with ECT02; Charleroi: An417 with An47 and the 3 Nubo with each other).

The equivalence demonstration is done for Angleur and Charleroi. The common periods, for which the capture rate of the devices is higher than 75%, includes respectively 11 and 13 days.

- For Angleur, only the Grimm with a W_{CM} = 11. 9% passes directly the equivalence test, the other sensors need to be calibrated. Thereafter, four additional devices sensors pass the test, namely the Sa005 with a W_{CM} = 22.2% (68.7% before calibration), the N032 with a W_{CM} = 24.4% (52.5% before calibration), the N053 with a W_{CM} = 23.4% (53.6% before calibration) and the N111 with a W_{CM} = 22.8% (50.1% before calibration). All these devices were built with an SPS30 sensor.
- For Charleroi, only the Grimm fails the equivalence test and needs a calibration to pass with a W_{CM} = 22% (66.5% before calibration). Devices based on both the Honeywell and the Sensirion sensors pass directly the test.

Table 3 summarizes the results of the demonstration of equivalence, including calibration equation. In order to complete the analysis, one can mention that the average and the standard deviation of the reference method in Angleur are 5.46 µg/m^3 and 3.47 µg/m^3, and in Charleroi 13.08 µg/m^3 and 6.45 µg/m^3.

In addition, a Bland-Altman plot [5] is drawn for each device in respect with the reference method (see Figure 3). This representation of method agreement is made by taking the mean of each couple "device and reference values" on the x-axis and its difference on the y-axis. Ideally, all points should be scattered as

closely as possible to 0 for the whole range of observed values. The bias corresponds to the mean error. The limits of agreement correspond to 1.96 times the standard deviation of the differences against the bias. For the measurement campaign in Charleroi, one can see in the upper left corner subplot that the bias of the Grimm is far from 0 with a value of -4.5 µg/m^3, that the repeatability (half the distance between the limits of agreement) is high with a value of 9.96 µg/m^3 and that a tendency of increasing gaps with increasing values is displayed; all these elements make the instrument, a priori, a not so good candidate for the equivalence. Nevertheless, the impact of calibration is also directly visible on the Bland-Altman plot, as shown in Figure 4. The limits of agreement are widely reduced and the bias is nearly equal to 0; the calibration equation y = bx explains the term "nearly". By using the calibration factors determined in 2010, the repeatability decreases from 4.98 µg/m^3 to 2.97 µg/m^3 and the Grimm fails the test of equivalence (27.5%).

		u	W_{CM} (%) BC	W_{CM} (%) AC	b	u_b	a	u_a	Correction
Angleur	**GRIMM**	1.37	11.9	-	-	-	-	-	-
	An310	1.87	74.8	76.4					
	An38	1.77	89.4	82.3					
	An411	1.66	99.9	27					
	An417	NA	NA	NA	NA	NA	NA	NA	NA
	An47	1.75	100.5	27.5					
	Sa003	NA	NA	NA	NA	NA	NA	NA	NA
	Sa005	1.22	68.7	22.2	0.5772	0.0817	2.0228	0.5217	$y_{i.cal}=(y_i-a)/b$
	N032	1.30	52.5	24.4	0.7275	0.0992	0.3300	0.6330	$y_{i.cal}=y_i/b$
	N053	1.37	53.6	23.4	0.7283	0.0975	0.1705	0.6220	$y_{i.cal}=y_i/b$
	N111	1.31	50.1	22.8	0.7472	0.0968	0.1270	0.6178	$y_{i.cal}=y_i/b$
	ECT02	1.56	64	36					
	ECT03	1.34	59.4	40.5					
Charleroi	**GRIMM**	3.62	66.5	22	1.2988	0.0909	0.5964	1.3155	$y_{i.cal}=y_i/b$
	An310	0.79	9.4	-	-	-	-	-	-
	An38	0.85	11.5	-	-	-	-	-	-
	An411	NA	NA	NA	NA	NA	NA	NA	NA
	An417	2.04	20.7	-	-	-	-	-	-
	An47	2.19	23.5	-	-	-	-	-	-
	Sa003	1.63	19.8	-	-	-	-	-	-
	Sa005	1.15	10.2	-	-	-	-	-	-
	N032	1.31	16.2	-	-	-	-	-	-
	N053	1.29	12.6	-	-	-	-	-	-
	N111	1.27	13.2	-	-	-	-	-	-
	ECT02	NA	NA	NA	NA	NA	NA	NA	NA
	ECT03	NA	NA	NA	NA	NA	NA	NA	NA

BC = before calibration ; AC = after calibration ; scope = W_{DQO}<25%

Table 3. Demonstration of equivalence for Angleur and Charleroi. Devices that pass the test are in green, devices that fail it in red.

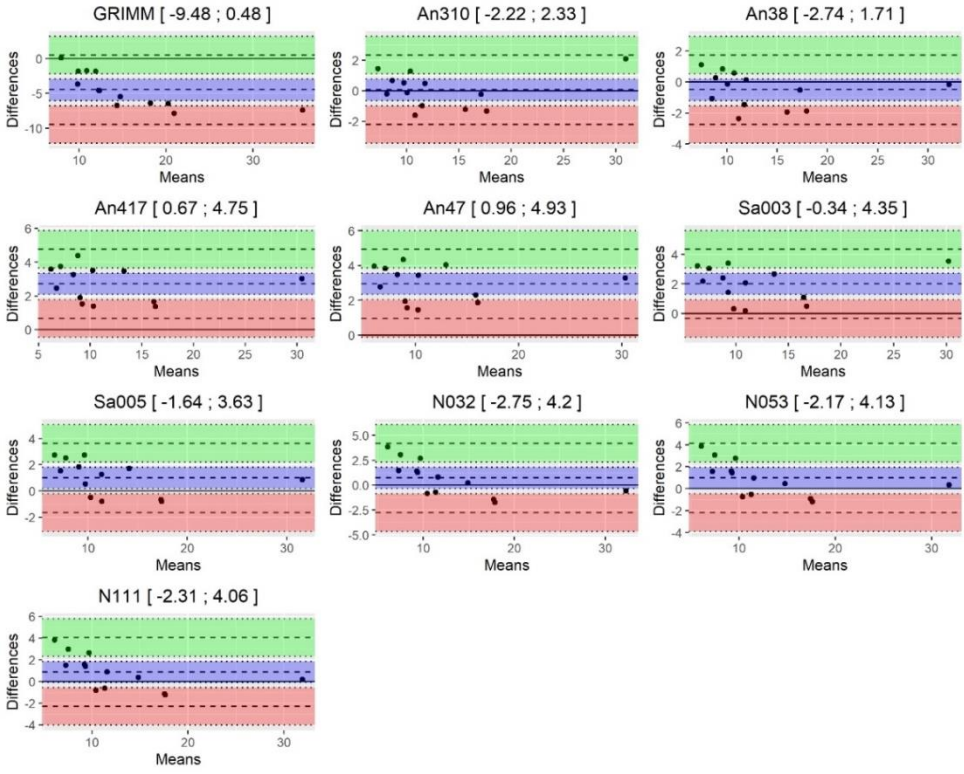


Fig. 3. Bland-Altman plots for the measurement campaign in Charleroi. The intermediate dashed line represents the bias between the method, both others the upper and the lower limit of agreement (their value is also in the title of each subplot). The green ribbon corresponds to the IC 95% on the upper L.A., the blue one on the bias and the red one on the lower L.A. The difference corresponds to "candidate-reference".

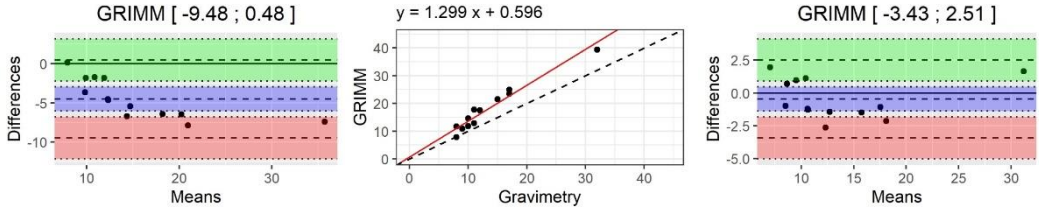


Fig. 4. Bland-Altman plots before (left) and after (right) calibration and the regression plot (middle) for the Grimm in Charleroi.

Table 4 includes different parameters extracted from the Bland-Altman analysis, the Spearman's and Lin's coefficients and the traditional MAE and RMSE.

From our two experiments one can see that Spearman's and Lin's coefficients are not perfectly correlated, meaning that a good rank correlation does not necessarily lead to a good concordance correlation and conversely, e.g. Grimm in Charleroi. According to a subjective classification in the literature [7], all sensors labeled "Excellent" or "Very good" (>0.95; [0.91-0.95], respectively) pass directly the equivalence test and those labeled "Poor" or "Mediocre" ([0.51-0.6]; [0.61-0.7], respectively) always fail the test even with a calibration. Unfortunately, the grey zone of "Satisfactory" and "Fairly Good" ([0.71-0.8]; [0.81-0.9], respectively) labels does not allow one to conclude anything.

As well, one cannot draw any conclusion based on both chosen error metrics. They usually depend too much on the concentration levels observed; they may be useful in relative terms but it has not been tested here.

	L.A. lower	L.A. upper	Bias	Accuracy	Repeat.	Slope	Spearman	Lin		MAE	RMSE
GRIMM	-4.05	1.09	-1.48	1.76	2.57	0.13	0.73	0.82	Fairly good	1.75	1.93
An310	-6.01	4.23	-0.89	2.4	5.12	0.59	0.71	0.55	Poor	2.40	2.64
An38	-5.33	4.95	-0.19	2.07	5.14	0.7	0.71	0.55	Poor	2.06	2.50
An411	-3.37	5.36	0.99	1.52	4.36	0.8	0.72	0.61	Mediocre	1.52	2.34
An417	-	-	-	-	-	-	-	-	-	-	-
An47	-3	5.63	1.31	1.68	4.31	0.77	0.63	0.59	Poor	1.68	2.47
Sa003	-	-	-	-	-	-	-	-	-	-	-
Sa005	-3.2	3.77	0.28	1.18	3.48	0.52	0.86	0.80	Fairly good	1.18	1.71
N032	-1.8	4.11	1.15	1.26	2.95	0.3	0.82	0.81	Fairly good	1.26	1.84
N053	-1.61	4.23	1.31	1.34	2.92	0.3	0.86	0.80	Fairly good	1.34	1.93
N111	-1.56	4.07	1.25	1.28	2.81	0.28	0.86	0.81	Fairly good	1.27	1.85
ECT02	-5.04	3.24	-0.9	2.04	4.14	0.54	0.71	0.69	Mediocre	2.04	2.20
ECT03	-3.18	4.27	0.54	1.38	3.72	0.39	0.63	0.78	Satisfactory	1.37	1.89
GRIMM	-9.48	0.48	-4.5	4.52	4.98	-0.26	0.95	0.78	Satisfactory	4.52	5.12
An310	-2.22	2.33	0.05	0.94	2.27	0.03	0.93	0.98	Excellent	0.94	1.11
An38	-2.7	1.71	-0.49	0.96	2.20	-0.03	0.94	0.98	Excellent	0.95	1.20
An411	-	-	-	-	-	-	-	-	-	-	-
An417	0.67	4.75	2.71	2.71	2.04	-0.02	0.93	0.90	Very good	2.71	2.88
An47	0.96	4.93	2.94	2.94	1.98	-0.01	0.95	0.88	Fairly good	2.94	3.09
Sa003	-0.34	4.35	2.00	2	2.34	0.01	0.94	0.93	Very good	2.00	2.31
Sa005	-1.64	3.63	0.99	1.43	2.63	-0.07	0.95	0.96	Excellent	1.42	1.63
N032	-2.75	4.2	0.72	1.54	3.47	-0.16	0.93	0.96	Excellent	1.54	1.85
N053	-2.17	4.13	0.98	1.5	3.15	-0.12	0.93	0.96	Excellent	1.49	1.83
N111	-2.31	4.06	0.87	1.5	3.18	-0.13	0.93	0.96	Excellent	1.46	1.78
ECT02	-	-	-	-	-	-	-	-	-	-	-
ECT03	-	-	-	-	-	-	-	-	-	-	-

Table 4. Limits of agreement (L.A.), bias, accuracy, coefficient of repeatability and slope of the Bland-Altman plots, Spearman's rank correlation coefficient, Lin's concordance correlation coefficient, Mean Absolute Error and Root-Mean-Square Error for Angleur (above) and Charleroi (below). Devices that pass the equivalence test directly are in green, devices that pass it after calibration in orange and devices that fail it in red.

4. Conclusions

In this limited experiment, some devices based on low-cost sensors display performance similar to a higher-end instrument with respect to the demonstration of equivalence methodology. However, it is worth

noticing that only two sites are investigated, that the range of measured values is relatively limited and that a rather short period of the year is covered.

All these parameters are performance indicators and could be used as-is or in combination to evaluate the metrological quality of a device. However, to find an alternative to the demonstration of equivalence without adding some subjective or site-dependent thresholds seems, on the mere basis of these two measurement campaigns, quite unlikely. The Bland-Altman plot provides an interesting visual inspection of the data set and seems promising; it will require some additional work to set the bias and L.A. values that could hopefully be used in all sites.

From these limited test results, it seems that the sensor performance really depends on the type of site investigated. One can assume that these differences are due to different calibration method for each manufacturer.

In the future, we will set up two additional measurement campaigns to evaluate the performance on a colder period and on two different sites, and evaluate the performance of the devices for the other parameters (CO, NO, NO_2 and O_3).

Acknowledgments

We would like to thank Didier Muck who helped us set up all instruments for the measurement campaign, Laurent Spanu for checking the implementation of parts of the R code, Robin Laruelle and Sébastien Fays for providing the Grimm data for the Charleroi experiment and Laurent Collard for the conception of the devices of ULiege. We are also grateful to AwAC for sharing and allowing the use of the measurements from the air quality monitoring network of Wallonia, ISSeP for supporting the OIE and Microcapteurs projects that respectively gave birth to the Antilope v.3 and 4, and to the Saïga, as well as the Walloon region for supporting the EcoCityTools project that gave birth to the ECT.

References

[1] Redon, N.; Spinelle, L. (2018): Premier essai national d'aptitude des micro-capteurs (EAµC) pour la surveillance de la qualité de l'air : synthèse des résultats. In: https://www.lcsqa.org/system/files/rapport/LCSQA2017-CILmicrocapteurs-synthese_resultats.pdf

[2] Fishbain, B. et al. (2016): An evaluation toolkit of air quality micro-sensing units. In: Science of The Total Environment 575, 639-648 http://dx.doi.org/10.1016/j.scitotenv.2016.09.061.

[3] EC Working Group (2010); Guide to the demonstration of equivalence of ambient air monitoring methods. In: https://ec.europa.eu/environment/air/quality/legislation/pdf/equivalence.pdf

[4] Bulot, F.M.J., Johnston, S.J., Basford, P.J. et al (2019): Long-term field comparison of multiple low-cost particulate matter sensors in an outdoor urban environment. In: Science Reports 9, 7497

[5] Feenstra, B. et al. (2019): Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site. In: Atmospheric Environment 216, 116946, https://doi.org/10.1016/j.atmosenv.2019.116946.

[6] Bland, J.M.; Altman, D.G. (1999): Measuring agreement in method comparison studies. In: Stat Methods Med Res. 8, 135-60.

[7] Patrik, B.L.; Stadler, A.; Schamp, S.; Koller, A.; Voracek, M.; Heinz, G.; Helbich, T.H. (2002): 3D versus 2D ultrasound: accuracy of volume measurement in human cadaver kidneys. Invest Radiol. 37, 489-95.

Intercomparison between IoT air quality monitoring devices for PM10 concentration estimations

Paulo G. Pinho[1,2], Sérgio M. Lopes[1,3], Marios Panourgias[4], Johnny Reis[5], Kostas Karatzas[4]

1. Introduction

Health problems associated with exposure to air pollutants such as particulates ($PM_{2,5}$ and PM_{10}), Ozone (O_3) and Nitrogen Dioxide (NO_2) are a major concern, (EEA, 2019). In 2017, concentrations of particulate matter (PM) continued to exceed the EU limit values and the WHO air quality (AQ) guidelines in large parts of Europe. For PM_{10}, levels above the EU daily limit value were registered at 22% of the reporting stations in 17 of the 28 EU Member States and in six other reporting countries (EEA, 2019).

In recent years, there has been a substantial development of technologies for measuring ambient air pollution. These developments have led to new AQ monitoring devices that include sensors and communication components with lower costs and energy requirements, thus leading to temporal and spatial representativeness improvement.

For PM, optical sensors mainly based on particle counting procedures with the aid of a light beam are utilised. PM concentration is further determined by signal processing (based on the separation of particle size as a function of beam interruption time) and by a algorithm, that considers particle density and counting efficiency amongst other parameters. The final concentration estimations depend on the operating conditions and the algorithm used for converting particle counts to concentration (Gozzi *et al.*, 2016, Johnson *et al.*, 2016; Williams *et al.*, 2014).

The present work compares the performance of three AQ monitoring devices in a field test made in Mediterranean conditions in Thessaloniki, Greece.

2. Materials and methods

The field tests were made as a part of an intercomparison exercise organised by the Environmental Informatics Research Group (EIRG) since 2018 in Thessaloniki, Greece. The goal of the exercise is to compare the performance of various low-cost air quality monitoring devices under field conditions, in a Mediterranean environment characterised by both high PM and gaseous pollutant levels, and also by a multiplicity of meteorological conditions that span from dry, hot summers to wet, cold, windy winters. For

[1] Center for Studies in Education, Technologies and Health (CI&DETS), Polytechnic of Viseu, Viseu, Portugal, ppaulo@estv.ipv.pt, slopes@estv.ipv.pt

[2] CITAB – Centre for the Research and Technology of Agro-Environmental and Biological Sciences, Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal

[3] ADAI/LAETA, Association for the Development of Industrial Aerodynamics, Coimbra, Portugal

[4] Environmental Informatics Research Group, School of Mechanical Engineering, Aristotle University, Thessaloniki Greece, mpanourg@meng.auth.gr, kkara@auth.gr

[5] Department of Environment and Planning & CESAM, University of Aveiro, 3810-193 Aveiro, Portugal, johnnydreis@ua.pt

this purpose, the location of the exercise was selected to be the rooftop of the 2-storey building (10 m height) of the School of Mechanical Engineering located in the 3rd of September street (40° 37' 37 N, 22° 57' 40 E), this being a typical city centre location capable of depicting traffic and other urban-related emissions.

Thessaloniki is the second most populated city in Greece where traffic and residential heating are the main urban air pollution sources with an additional source from the extended industrial area located at the west/northwest of the city. Although high levels of air pollution may appear in various locations and for various pollutants, PM has been the most pronounced pollutant in the area (EEA, 2006).

The air quality monitoring devices employed in the field tests include: (a) Monitar SmartAirsense - MSAS (manufacturer: MonitarSense, Portugal), monitoring PM_{10} and $PM_{2,5}$ (via a PM2105-M laser particle sensor from Cubic Sensor and Instrument Co. Ltd) as well asCO, NO_2 and O_3; (b) Dust Sentry PM10 - DS-PM10 (manufacturer: Aeroqual Ltd, New Zeland), monitoring PM_{10} and $PM_{2,5}$; and (c) PurpleAir PA-II (manufacturer: PurpleAir Llc, USA), monitoring PM_{10} and $PM_{2,5}$. MSAS is an AQ monitoring system supporting remote operation and it includes an optical sensor for PM and electrochemical sensors for O_x, NO_2 and CO. These sensors are integrated, and the signal is optimised by MonitarSense algorithms who also performs sensor calibration. The calibration is done against reference methods, defined in Directive 2008/50/EC and by an accredited test laboratory, in factory conditions and before being shipped to the field. The device includes temperature control, sample pump and volumetric flow controller in two channels, one for PM and one for gaseous pollutants.

DS-PM10 is based on a light scattering nephelometer that has a sharp cut cyclone and an on-board temperature sensor to correct for thermal drift, sheath air filter to keep the optics clean, automatic baseline drift correction, and a fibre optic span to enable a check of the optical components. It also includes omni-directional heated sampler inlet, sample pump and volumetric flow controller. PA-II is a compact device that employs the Plantower PMS5003 optical particle counter. These sensors come with a built-in miniature fan to secure the necessary inflow, count suspended particles in various size ranges up to 10μm. PMS5003 sensors come calibrated from factory. PurpleAir PA-II uses two sensor units attached to each other and placed in the same shelter.

In terms of data analysis, we studied the whole time series and subsets of the available data, for reasons of missing data (September for the DS-PM10 device) as well as in order to acquire a deeper knowledge of the behaviour of the devices in different time periods of the year. It was performed a standard least square regression with the calculation of the determination coefficient as a goodness of fit statistic. It should be mentioned that due to the distance between the location of the nearest reference (official) PM10 measurements and the location of our intercommunication exercise (more than 1 km), it was not considered to be appropriate to include aforementioned data in the intercommunication.

In order to further investigate the relationships and patterns of behaviour of the monitored parameters and therefore of the relevant devices, we employed a computational intelligence method that can account for non-linear relationships and does not require supervised training, *i.e.* Self Organizing Maps (SOM) (Kohonen, 1982). This is a method based on neural networks that is capable of mapping a multidimensional feature space (here the various parameters being monitored) to a type of map (here a 2D lattice), that reflects via its topological characteristics (distances and neighbourhoods of points) the relationships and patterns of the initial features. The training procedure of the SOM is based on the initialisation of a two-dimensional array of (initially) randomly weighted neurons. All data points (vectors of the feature space) are passed through the neural network and are matched with a winning neuron, causing the network topology to adjust and eventually form clusters of similar attributes, while weights are updated to better fit into the process.

In this way, the final map (created for each one of the parameters of the feature space) represents "neighbourhoods" of features which are topologically compared. The latter depends on the unified distance matrix (U-matrix), which is commonly used for SOM visualization, and represents the Euclidean distance between neighbouring neurons which is actually an expression of the relationship ("similarity") between those neurons and consequently of the features they represent (Ultsch and Siemon, 1990). Here, we used the SOM Toolbox for Matlab (http://www.cis.hut.fi/somtoolbox/) that offers a stable and commonly used implementation of the method

3. Results

The intercomparison time interval is from 4th of February 2019 to 30th of September 2019, *i.e.* for a total of 239 days.

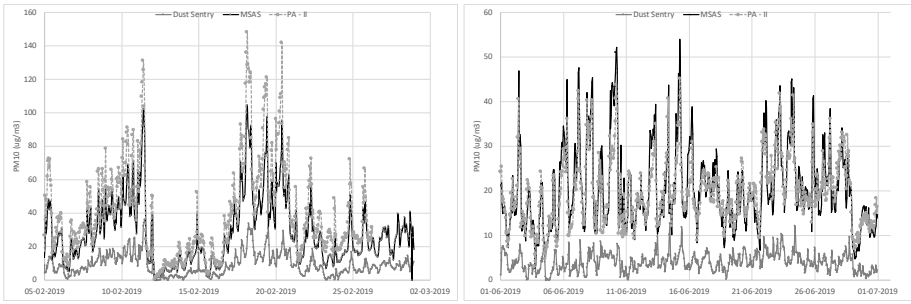
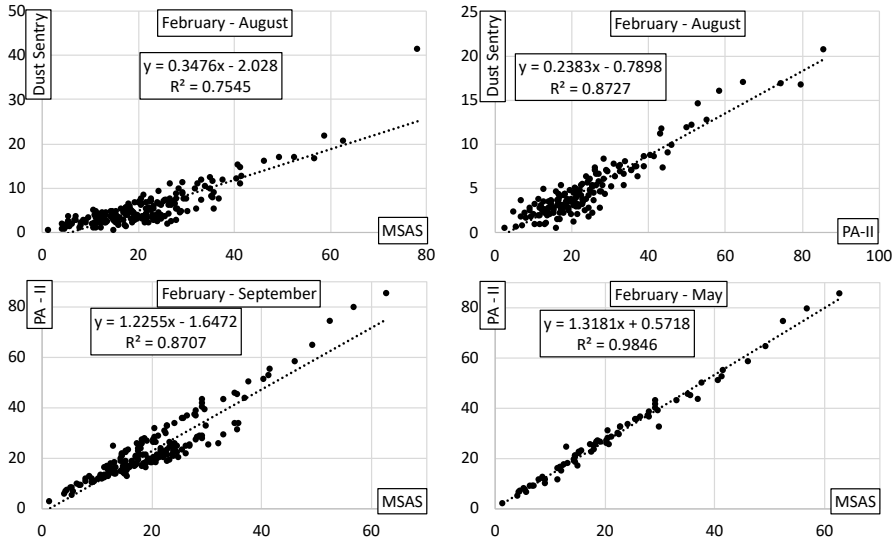


Fig. 1. Hourly PM10 measurements by air monitoring stations during February 2019 (up) and June 2019 (down).



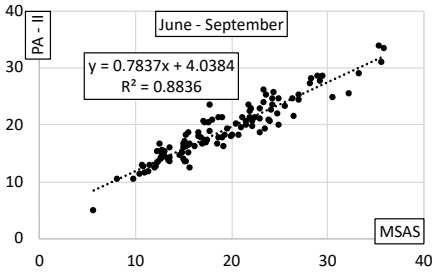


Fig. 2. Correlation between PM10 day measurements ($\mu g/m^3$) by air monitoring stations.

For both the typical winter as well as typical summer months, the DS-PM10 provides values that are lower than the ones provided by the MSAS and the PA-II devices (see Figure 1). Nevertheless, the diurnal pattern of the DS-PM10 seems to be in line with the diurnal patterns for the other two devices. The coefficient of determination between the DS-PM10 and the MSAS (0.75) is lower than the coefficient of the same instrument with the PA-II (0.87). On the other hand, the MSAS and the PA-II devices demonstrate a clear linear relationship for the period February to May 2019 (R^2=0.98), which still is the case for the summer period (June to September), yet with a slightly lower relationship (R2=0.88). This deviation and the linearity of the relationship between the two instruments leads to a value of R^2=0.87 for the whole time period of study.

The DS-PM10 measurements were adjusted making use of their correlation with MSAS (Figure 3 for June 2019). The MSAS was used for adjustment since the MSAS was calibrated, before the field test, with the reference method (gravimetric) with a very good relation. The adjusted data are well in line with MSAS and PA-II, yet this does not provide additional insight on whether the DS-PM10 underestimated PM_{10} concentrations or whether the two other devices overestimated them.

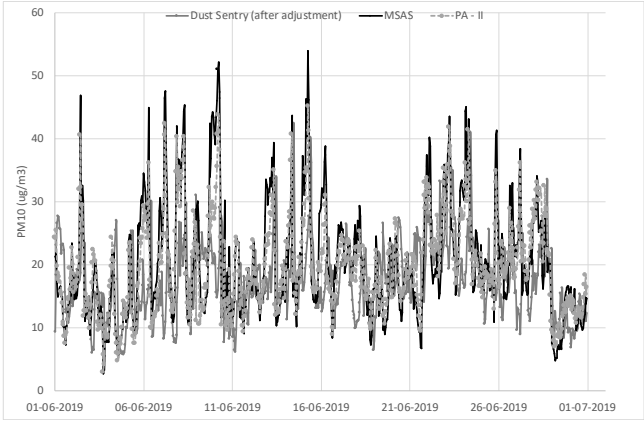


Fig. 3. Hourly PM10 measurements ($\mu g/m^3$) by air monitoring stations during june 2019. PM10 concentrations from DS-PM10 were adjusted with correlation obtained with MSAS.

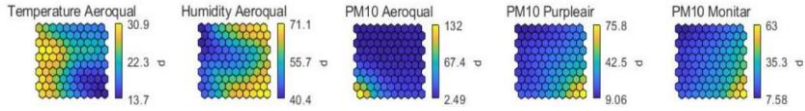


Fig. 4. The Self Organizing Maps for PM_{10}, temperature and humidity for the time period of the field intercomparison.

The SOMs for PM_{10}, temperature and humidity reveal that there is an area of PM_{10} values provided by PA-II and MSAS (lower-right in the relevant maps, see Figure 4) that completely match each other and coincide with high relative humidity (RH) values. This suggests that both instruments have a similar pattern of behaviour, which seems to be influenced from RH. On the other hand, the PM_{10} measurements coming from DS-PM10 present a different pattern, with the highest values coinciding with high temperatures.

Overall, the intercomparison for PM_{10} shows a good linear relation between all the tested devices.

The results from MSAS and PA-II are correlated in two different sets: between February 4 to May 7 and from May 23 till the end of September. The DS-PM10 started giving abnormal results (*i.e.* too high values) after August 19 and that data was not considered valid. PM_{10} measurements by MSAS and PA-II were in expected range when compared with the mean PM_{10} levels reported in Thessaloniki while the ones coming from the DS-PM10 were in the lower part of the range of PM_{10} levels measured in other locations of the city (EEA, 2020). It should nevertheless be noted that the DS-PM10 should run zero calibration every month (not done in our case), and flow check, leak check and filter replacements at frequencies recommended by the manufacturer, while all instruments should be calibrated against a reference method.

An additional limitation is the inherent sensitivity of the optical counters to high humidity levels, that may affect mean size (and therefore overall mass estimation) of the simple sensors used by MSAS and PA-II, due to the inherent hygroscopicity of PM.

Coming to the patterns of the PM_{10} concentrations, the SOMs analysis revealed that the PA-II demonstrates a similar pattern with the MSAS measurements, while both seem to be more influenced by relative humidity. The DS-PM10 instrument presented with the biggest differences in terms of its behaviour in comparison to the other two devices.

4. Conclusions

All devices use an optical method for measure PM_{10}, yet supported by additional flow and measurement technologies. All instruments demonstrated a linear relationship as well as good determination coefficients to each other. PM_{10} reported by the DS-PM10 were lower in comparison with MSAS and PA-II. The latter two devices share an optical counter sensor for PM_{10} of very similar technology while DS-PM10 uses a nephelometer for PM_{10} estimations, which nevertheless was not maintained in full compliance with the manufacturer's guidelines in the case of the specific field test.

Overall, AQ monitoring devices that do not belong to the reference instrument category have been proven to be capable of estimating the level and the profile of PM_{10} in urban environments. The optical counting sensors depict the temporal profile of the PM_{10} in a similar way, demonstrate a satisfactory linear relationship and a good determination coefficient, while they are characterised by differences in the reported absolute values. This means that more research is required in order to calibrate such instruments against reference methods, and especially in order to computationally improve non-reference devices (Borrego et al., 2018), to render them appropriate for urban air quality monitoring, in line with the uncertainty requirements of the clean air for Europe directive. Overall, it is necessary to further develop the

devices and to implement validation and calibration protocols that will guarantee the quality of results (MacDonell *et al.*, 2013; Snyder *et al.*, 2013; Spinelle *et al.*, 2013; Jovašević-Stojanovića *et al.*, 2015).

Acknowledgements

The authors acknowledge kartECO for providing access to the DS-PM10 data as well as Getmap Geospatial Enabling Technologies for providing access to the PurpleAir data in the frame of the Public Participation CITY project.

References

[1] Borrego C., Costa A.M., Ginja J., Amorim M., Karatzas K., Sioumis Th., Katsifarakis N., Konstantinidis K., De Vito S., Esposito E., Smith P., André N., Gérard P., Francis L.A., Castell N., Viana M., Minguillón M.C., Reimringen W., Otjes R.P., v.Sicard O., Pohle R., Elen B., Suriano D., Pfister V., Prato M., Dipinto S., Penza M.. Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise– part II, Atmospheric Environment 193, 127-142. (2018).

[2] European Environmental Agency (EEA). Air Pollution at Street Level in European Cities. EEA Technical Report No 1/2006. (2016).

[3] European Environmental Agency (EEA). Air quality in Europe. Report. EEA Report n.º 10/2019. (2019).

[4] European Environmental Agency (EEA). Air Quality Time Series, http://discomap.eea.europa.eu/map/fme/AirQualityExport.htm. (2020).

[5] Jovašević-Stojanovića, M., Bartonovab, A., Topalovićc, D., Lazovića, I., Pokrićd, B., Ristovskie, Z. On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter. Environmental Pollution, Vo. 206, 696–704. (2015).

[6] Kohonen, Teuvo. Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics 43 (1), pp. 59–69. (1982).

[7] MacDonell M., Raymond M., Wyker D., Finster M., Chang Y., Raymond T., Temple B., Scofield M. Research and Development Highlights: Mobile Sensors and Applications for Air Pollutants. Argonne National Laboratory Environmental Science Division (EVS) Argonne, IL, EPA/600/R-14/051. (2013).

[8] Snyder, E., P. Solomon, M. MacDonell, R. Williams, E. Thoma, D. Vallano, M. Raymond, D. Olson. Next generation air monitoring-a review of portable air pollution sensors. Presented at 2013 32nd annual AAAR, Portland, OR, September 30. (2013).

[9] Spinelle L., Aleixandre M., Gerboles M. Protocol of evaluation and calibration of low-cost gas sensors for the monitoring of air pollution, EUR 26112. (2013).

[10] Spinelle L., Gerboles M., Villani M. G., Aleixandre M., Bonavitacola F. Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. Sensors and Actuators B: Chemical, Volume 215, Pages 249-257. (2015)

[11] Spinelle L., Gerboles M., Villani M. G., Aleixandre M., Bonavitacola F. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2. Sensors and Actuators B: Chemical, Volume 238, Pages 706-715. (2017).

[12] Ultsch A. and Siemon H.P." Kohonen's Self Organizing Feature Maps for Ex-ploratory Data Analysis" Proceedings of International Neural Networks Conference (INNC) (1990), pp. 305-308. (1990).

ECOSense and its preliminary findings: Collection and analysis of bicycle sensor data

Johannes Schering[1], Christian Janßen[2], René Kessler[3], Viktor Dmitryev[4], Jorge Marx Gómez[5], Christian Stehno[6], Kyra Pelzner[7], Ronald Bankowsky[8], Roland Hentschel[9]

1. Introduction and Motivation

Fig. 1. Sensor with open enclosure (photograph by CoSynth).

The bicycle is increasingly perceived as an attractive alternative to the private car and especially at pandemic times to keep social distance. Facing the increasing debates about climate protection and mobility problems, municipalities are increasingly recognizing the potential of cycling and want to promote a more frequent bicycle use. During the Corona pandemia more and more cities as Berlin for example implement pop-up bike lanes as short time solution to adjust the traffic infrastructure to the demand of growing numbers of cyclists [1]. However, the prerequisite for perspective improvements of cycling conditions regarding an attractive and inviting bicycle infrastructure, which includes more direct connections from the surrounding area to the cities, is the availability of appropriate data. It is a big problem that there is no or only too little suitable data on bicycle traffic available that could be used by traffic or city planners for the further expansion of the infrastructure [2]. So far, data-centered approaches regarding to cycling have been primarily based on information generated by smartphones. The existing database primarily concentrates on leisure traffic and provides only little information about everyday bicycle use. In addition, there is no data available on the condition of cycling paths (e.g. surfaces) as smartphone applications mainly gather data on

[1] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, johannes.schering@uni-oldenburg.de
[2] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, christian.janssen@uni-oldenburg.de
[3] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, rene.kessler@uni-oldenburg.de
[4] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, viktor.dmitryev@uni-oldenburg.de
[5] University of Oldenburg, Department of Computing Science, Business Informatics (Very Large Business Applications VLBA), Oldenburg, Germany, jorge.marx.gomez@uni-oldenburg.de
[6] CoSynth GmbH & Co. KG, Oldenburg, Germany, stehno@cosynth.com
[7] baron mobility service GmbH, Oldenburg, Germany, kyra.pelzner@baronmobil.com
[8] baron mobility service GmbH, Oldenburg, Germany, ronald.bankowsky@baronmobil.com
[9] City of Oldenburg, Department of Economy Promotion, Oldenburg, Germany, roland.hentschel@stadt-oldenburg.de

bicycle use (tracking). As a consequence, existing research about the digitalization of cycling is mainly based on smartphone generated data (see the project Movebis[10]). This leads to a research gap in the field of *smart cycling* as different target groups are not well represented (e.g. elderly who do not use a smartphone) and data on the shape of the bicycle infrastructure (e.g. vibration of the bicycle when cycling on the bicycle path) is not available.

The project ECOSense[11] filled out this research gap by a new sensor based approach under active citizen participation. Project partner CoSynth developed and tested a sensor platform that collects various parameters of cycling sensor data (position, speed, vibration, environment) on everyday bicycle use. The newly generated and refined data sets lead to an improved information base about cycle use in daily life (e.g. commuting to work or place of education). The implementation of the measurement technology enables decision-makers from municipal and traffic planning to understand the specific needs of cyclists better than before and enables the adjustment of bicycle infrastructures to specific demands of cyclists. ECOSense (project number VB18F1030) was funded by the Federal Ministry of Transport and Digital Infrastructure (BMVI) of Germany as part of the funding program Modernity Fund ("mFUND")[12] with a funding amount of 97.272€ and includes the three project partners mein-dienstrad.de (baron mobility service GmbH, lead partner), CoSynth GmbH & Co. KG and the University of Oldenburg - Department of Business Informatics, Very Large Business Applications (VLBA) under the lead of Professor Dr.-Ing. habil. Jorge Marx Gómez. According to the project plan and the approval the feasibility study was running for one year. In spring 2020 the lifetime of the project was increased for another three months because of the spread of the Corona virus and related delays. In total, the 15-month feasibility study was running from June 2019 until the end of August 2020.

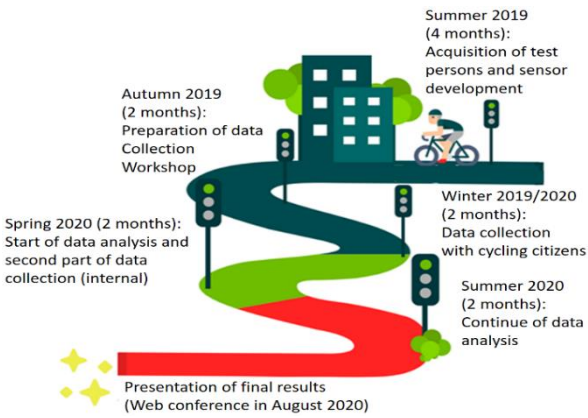


Fig. 2. Overview of the project scope and processes.

Figure 2 shows the working processes and the time scale of the project. ECOSense started in Summer 2019 with the sensor development (section 2) and the acquisition of cycling citizens as active sensor users (section 3, Data collection methodology). In autumn 2019 the participants were selected and informed. A workshop with traffic experts were conducted to evaluate the ECOSense sensor approach and discuss about new research ideas based on bicycle sensor applications (section 3, Stakeholder engagement, communication). Between November 2019 and February 2020 a group of 200 citizens collected bicycle data (about two months per participant). Because of the Corona contact restrictions only project members and surrounding people were part of the second round of data

[10] https://www.movebis.org/
[11] https://ecosense.mein-dienstrad.de/
[12] https://www.bmvi.de/DE/Themen/Digitales/mFund/Ueberblick/ueberblick.html

collection in spring and summer 2020 (section 3, Data collection methodology). Based on the first round of data collection the University started the work on the data analysis part parallelly (section 4, Preliminary results of the data analysis). The preliminary results of the data analysis based on both rounds of data collection were presented as part of an online based web conference in August 2020 (section 3, Stakeholder engagement, communication).

2. Sensor development

As part of the first iteration step of the project, the partners discussed the requirements on suitable measurement technologies. Therefore, parameters of the bicycle sensor were specified such as brightness, temperature, air pressure, NO_x or fine dust. The two latter mentioned environmental parameters were not integrated because of cost reasons and high demand for energy supply. Figure 1 shows the sensor as used for measuring with battery and housing, but with removed front cover. The collected environmental data includes air pressure, temperature and humidity. The sensor is measuring position, time and acceleration using a GPS module. As key data of highest interest the vibration of the bicycle due to surface roughness is measured which makes the sensor platform an innovative and unique approach for cycling measurement. The gathered sensor data enables new analyses about the cycling conditions of a city or region. At the final development stage of the project the cycling data is transferred manually offline instead of online in real time. Figure 3 shows the procedure of the data transfer.

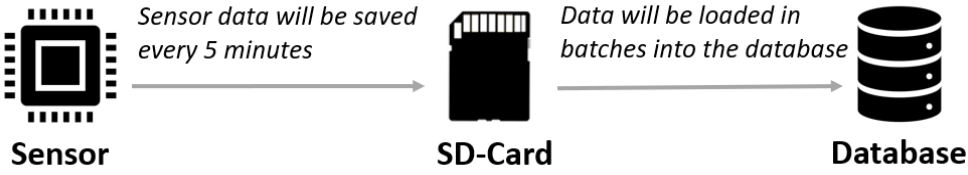


Fig. 3. Process of data transfer.

The sensor data is saved encrypted on an SD (Secure Digital) card inside the sensor box. External access (e.g. by WLAN) is prohibited for security and energy reasons. As the current sensor may collect data for about 100 hours, the participants were not provided with charging systems. Depending on bicycle use, the batteries are replaced every four to eight weeks. The sensor housing was equipped with hook-and-loop mounting strips to mount the sensor on the frame of the bicycle. Figure 4 shows the final design of the sensor platform with closed cover. The sensors were attached at various positions of the bicycle frame (saddle tube, handlebar etc.). Regarding data quality no significant differences were identified.

Fig. 4. Sensor unit developed by CoSynth, attached to a bicycle (photograph by University of Oldenburg).

3. Experimental Plan

Data collection methodology

Project partner mein-dienstrad.de (baron mobility service GmbH) was involved as an application partner in the area of bicycle leasing with access to many bicycle friendly organizations and was responsible for distributing the 200 sensors to everyday cyclists in the Oldenburg area over a period of several months. The lead partner organized the sensor supply to the interested participants.

ECOSense attracted high attention by the general public in the Oldenburg area in Lower Saxony/Germany. A call-up for participation was started on different communication channels. The local newspaper Nordwest-Zeitung (NWZ) published several articles about ECOSense [3]. The German Bicycle Club ADFC, district association Oldenburg, sent a call for participation to its members by email. Therefore, the acquisition of project participants was completed after a short time period as many citizens were interested to participate in the project. Just a few days after starting the application phase on the official website, several hundreds of people registered for participation. The ECOSense project proved that citizens are willing to participate in citizen science projects and to provide some personal data if the personal benefit is clear and easy to understand. Many participants stated that their main motivation for participation is their hope to contribute to the improvement of the bicycle infrastructure in mid- and long term perspective.

Another interesting result is that new user groups of cyclists are reached by sensor applications. Compared to bicycle apps which are mainly used by younger, sportive people during spare time [4] elderly

people are more attracted to participate (8,41 percent of the participants were older than 66 years). Nearly 50 percent of the people are older than 46 years. The amount of people older than 66 years was higher than the amount of young adults (18-25 years). That means that compared to the usage of bicycle smartphone applications, seniors are quite well represented as students are quite underrepresented. In the case of ECOSense, more than 520 citizens from all districts of the city of Oldenburg and its suburbs registered to participate in the project. About 59 percent of the cyclists were male, 41 percent female.

18-25 years	26-35 years	36-45 years	46-55 years	56-65 years	66 years or older
38 participants (7,26%)	131 participants (25,04%)	95 participants (18,16%)	128 participants (24,47%)	87 participants (16,63%)	44 participants (8,41%)

Table 1. Age structure of the participants (provided by mein-dienstrad.de).

The results of the age distribution of the sensor users can be confirmed by experiences made in foreign countries. With an high emphasis on citizen participation the Sniffer Bike project[13] from the Netherlands follows a research approach which is very similar to ECOSense. Besides to the position, acceleration and speed levels the more than 500 bicycle sensors are gathering different types of environmental data (e.g. fine dust and temperature) in cities all over the Netherlands. According to the Province of Utrecht which is strongly engaged in the Sniffer Bike project demographic details of the participants were not compiled. However, evaluating by the distribution of the participants in several personal meetings and online interaction (e.g. email, web conference), middle aged and elderly people are likely to be overrepresented. The distribution of male and female participants is nearly equal. Also people with less experiences in technology driven approaches do participate actively in the project.

In the first round of citizen involvement in the data collection as part of the ECOSense project between November 2019 and February 2020 about 200 people from the Oldenburg area participated. Each person cycled with the sensor system for about two months. The original idea according to the project plan was to start a second round of data collection with Oldenburg based citizens in March and April 2020 after gathering the bicycle data of the first implementation phase. However, due to the nationwide shutdown in March 2020 in consequence of the Corona virus and the following contact restrictions it was not possible to organize a second large scale round of citizen involvement. Instead, a series of short routes were measured in order to improve the data analysis based on the specially chosen tracks.

Stakeholder engagement, communication

Besides the citizens many external stakeholders were involved in the course of the project. At a workshop held with city planners, traffic engineers, municipalities and researchers in October 2019 at Schlaues Haus, Oldenburg, the approach of the ECOSense project was discussed and further research topics were identified. In two workshop formats the idea of the research approach was critically evaluated by externals and ideas for further bicycle sensor applications were collected. The workshop with more than 40 domain experts showed that collecting cycling data with sensors is a relevant and interesting approach for potential later

[13] https://civity.nl/en/products-solutions/sniffer-bike/

users. Experts in traffic planning are seeking for sensor data to learn more about the quality of bicycle infrastructure and the specific demands of cyclists. The experts made clear that there is special interest in key factors such as parking times and time losses during cycling due to parked cars or traffic lights which are optimized for motorized traffic. Demands of other user groups such as pupils or elderly on alternative sensor applications were evaluated.

To disseminate the approach and results of ECOSense, the project was presented at different mFUND accompanied research formats (expert exchange on cycling data at the Federal Ministry of Traffic and Digital Infrastructure in Berlin, online web conference) and at the mFUND conference as part of the session "Detection and monitoring of traffic infrastructures" in Berlin. ECOSense was part of the German bicycle municipal conference (Fahrradkommunalkonferenz) in Wittenberg/Lutherstadt. About 270 stakeholders from municipalities all over Germany participated and were informed about the Oldenburg based sensor approach. ECOSense was presented at the Datacycle Meetup Berlin to data experts, bicycle activists and students. A project interview with focus on the main goals, target groups, challenges and first results was published on the BMVI website [5]. In August 2020 the project results were presented as part of a web conference. About 50 people (mainly citizens and traffic experts) participated and asked related questions to the project.

4. Preliminary results of the data analysis

Through the described collection rounds 26 GB of data consisting of route selection, acceleration, vibration and environment was collected. More than 270 sensors were applied to the bicycles of participating citizens and project employees. As a whole 13.838 bicycle trips were recorded. The database was anonymized to comply with data protection regulations. OpenStreetMap map material and data were used to visualize the adoptions of the project. As a scientific method for the data analysis CRISP-DM (Cross Industrial Standard Process for Data Mining) [6] with its various process steps has been conducted. The analysis of this bicycle sensor data research which is still work in progress focuses on vibration, route selection and weather conditions.

Vibration - Analysis of road quality (e.g. data patterns of surfaces, potholes, curbsides)

To analyse the road quality, the sensor measures vibrations occurring while cycling. For this purpose, primarily data of the acceleration sensor is used. The accelerometer measures the acceleration in three axes, of which the vibration ocurring on the bicycle can be calculated. As part of the data analysis significant events were collected and visualized as markers on a map. The markers on the map show spots with major vibrations. Green and yellow markers are aggregations of single blue markers and denote the total amount of vibrations nearby (Figure 5, left side, the right side shows a higher zoom level). As a future step possible causes of the vibration e.g. potholes, curbsides or dropped kerbs could be identified.

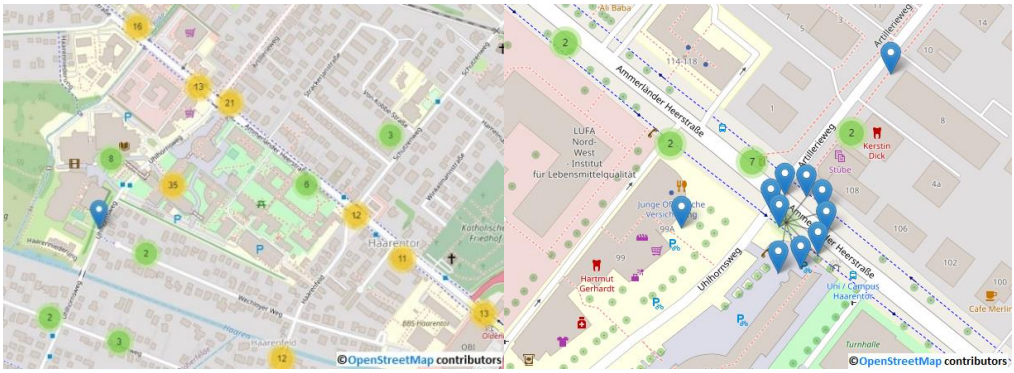


Fig. 5. Intensity of vibration during sensor measurements in Oldenburg (University of Oldenburg).

Route selection

During the analysis of the route selection of participants, the project determined that choices of different routes are based on the infrastructure. As shown in Figure 6, most of the cyclists prefer main roads with cycling paths. The main bicycle traffic volume is concentrated in this area. Further interpretations of the selection decisions will be carried out through pending surveys.

Fig. 6. Heat map of bicycle route selection in Oldenburg (University of Oldenburg).

Weather conditions

As part of the data analysis of environmental conditions, weather conditions were mainly focused in order to interpret the cycling behaviour under different weather situations. Therefore the database of meteostat[14]

[14] https://meteostat.net/de

(based on data of the German Weather Service Deutscher Wetterdienst DWD) was used and combined with the gathered data sets into data frames. It was found out that bad weather conditions shorten the average time and distance per run. In this specific case, participants cycled one kilometer less during bad weather conditions.

Good weather conditions:
Average time per run in minutes: 18.262878787878787
Average distance per run in kilometers: 5.493444165046406

Bad weather conditions:
Average time per run in minutes: 15.619402985074627
Average distance per run in kilometers: 4.49791445523026

A second analysis focused on weather conditions before the start of a trip which showed a clear trend. Bad weather conditions shorten the distance up to one kilometer as well. The average speed is slower by two kilometers per hour.

Good weather conditions before start:
Average time per run in minutes: 17.82063492063492
Average distance per run in kilometers: 5.382240374693439

Bad weather conditions before start:
Average time per run in minutes: 15.649333333333335
Average distance per run in kilometers: 4.392962313634

5. Conclusion and future work

As part of the mFUND funding program all projects are requested to publish open data sets at the open data portal of the Federal Ministry of Transport and Digital Infrastructure, the mCLOUD[15]. This is made in order to increase the amount of available mobility data for projects, applications and smart services to make mobility smarter, safer, more convenient and environmental friendly. As a result of ECOSense anonymized, sanitized and preprocessed sensor data will be published at the open data portal.

Another important result of the project is that participating citizens are very interested in their personal cycling data and in the conditions of the bicycle infrastructure (e.g. the cycle paths of the cycling network). When providing back the sensors to the project many participants asked whether and where they could visualize their personal bicycle trips. The project learned that there is a high demand for an open accessible dashboard which visualizes bicycle trips and conditions of bicycle paths. According to the users a dashboard would increase participation in further sensor projects. The Sniffer Bike project from the Netherlands mentioned above gives a good example how to publish the bicycle trips to the general public. The anonymized data is published and visualized on an open accessible dashboard on a daily basis.[16]

Further measurements based on the ECOSense bicycle sensors will especially focus on the choice of cycle paths and on surface qualities. Therefore, data patterns of surfaces, potholes, curbsides, dropped kerb etc. need to be identified in detail. The vibration measurements will be further analysed to learn more about different vibration levels and how to interpret these. The measurements need to be improved to deduce

[15] https://www.mcloud.de/
[16] https://dashboard.dataplatform.nl/sodaq/v2/groene_fietsroutes.html

further information on the conditions for cycling. This totally new information could be the basis for the assessment of the quality of bicycle lanes.

As part of the evaluation the participants of ECOSense are asked for their personal feedback in an ongoing online survey. The questions are mainly focusing on the motivation for participating in the project and on new ideas for further bicycle sensor projects. The results of the survey will have influence on further data analysis of the data base generated by ECOSense. The overall final results of the project were presented to the general public as part of a web conference. In addition, a project brochure with expert interviews, further information about the citizen involvement etc. will be published to disseminate the final results.

References

[1] Ensink, B. (2020). Temporäre Einrichtung und Erweiterung von Radverkehrsanlagen. In 10 Tagen mehr Platz für das Rad in der Stadt. https://www.mobycon.nl/wp-content/uploads/2020/04/6796_Kreuzberg_Handbuch-V4.pdf. In: Mobycon. Uploaded 2020. Accessed 17th of July 2020.

[2] Monheim, H.; Muschwitz, C.; Reimann, J.; Thesen, V.; Grade, A.; Kritzinger, S.; Rikus, S.; Leckel, S.; Gutberlet, T. (2016). Grundlagenuntersuchung zur Situation des Radverkehrs in Deutschland – Kurzdarstellung der Forschungsergebnisse. Study of raumkom (Institut für Raumentwicklung und Kommunikation) and Prognos AG (Europäisches Zentrum für Wirtschaftsforschung) on behalf of the Federal Ministry of Traffic and Digital Infrastructure. (FE-Nr. 70.0895/13). Trier, Berlin. http://mobilitaet21.de/wp-content/uploads/2018/02/700895_GrundlagenuntersuchungRadverkehr_Kurzdarstellung.pdf. Uploaded 2016. Accessed 22nd of July 2020.

[3] Kiefer, C. (2019). Sensoren zeigen Schwachstellen beim Radfahren. https://www.nwzonline.de/plus/oldenburg-testfahrer-in-oldenburg-gesucht-sensoren-zeigen-schwachstellen-beim-radfahren_a_50,6,132907604.html. In: Nordwest-Zeitung. Published in the Nordwest-Zeitung at the 7th of October 2019. Accessed 17th of July 2020.

[4] Francke, A. (2016): Auswertung der Befragung von Strava-NutzerInnen in Deutschland. In: TU Dresden. https://tu-dresden.de/bu/verkehr/ivs/voeko/ressourcen/dateien/forschung/nrvp_gps/vortrag-angela-francke-strava-nutzer?lang=de. Uploaded 2016. Accessed 1st of July 2020.

[5] Schering, J. (2019). mFUND Projekte im Porträt: 7 Fragen an ECOSense. https://www.bmvi.de/SharedDocs/DE/Anlage/DG/mFUND/siegen-fragen-an-ecosense.pdf?__blob=publicationFile. In: Federal Ministry of Traffic and Digital Infrastructure. Uploaded 27th of February 2020. Accessed 17th of July 2020.

[6] Wiemer, H.; Drowatzky, L.; Ihlenfeld, S. (2019): Data Mining Methodology for Engineering Applications (DMME)—A Holistic Extension to the CRISP-DM Model. In: Applied sciences, 01 June 2019, Vol.9(12), p.2407.

Towards a robust ensemble modelling approach to improve Low-Cost Air Quality Sensors performance

Theodosios Kassandros[1], Kostas Karatzas[2]

1. Introduction

Systematic monitoring of air quality has been carried out on the basis of a small number of stations in large urban areas, while air quality forecasting is carried out only as a research activity (as opposed to other environmental forecasts such as weather or wave forecasting). The KASTOM project is developing a versatile and flexible urban air quality monitoring and forecasting system by deploying an IoT-oriented network of low-cost air quality sensor nodes (LCAQSN), while in parallel developing a state-of-the-art emission modelling module combined with state-of-the-art three-dimensional air quality models. As the final KASTOM product includes detailed air quality estimates at an urban level, innovative data fusion methods as well as methods for the improvement of the performance of LCAQSN come into play. Recent studies indicate that Machine Learning (ML) may significantly improve the performance of air quality sensor nodes reducing the impact of cross-sensitivity issues as well as measurement uncertainty [1,2,3]. Aim of this study is to develop a pipeline of computational procedures (feature engineering, feature selection, ML algorithms, ensemble approach, validation), which is best suited for improving the performance of the aforementioned sensor nodes for particulate matter of 10 microns or less in aerodynamic diameter (PM10), in the frame of the KASTOM project.

2. Materials and Methods

2.1. Materials

Area of study

The Greater Thessaloniki Area (GTA) is the second largest urban agglomeration in Greece. Located in the Central Macedonia region, it hosts more than 1,100,000 inhabitants. The city of Thessaloniki is spread between the sea coast of Thermaikos Gulf (to its south and southwest), and the slopes of the Chortiatis Mountain (to the southeast) and the Seich-Sou forest to the northeast. The local climate is Mediterranean with hot, dry summers and mild, wet winters (Csa climate type according to the Koppen-Geiger climate classification (https://webmap.ornl.gov/ogc/dataset.jsp?ds_id=10012). Thessaloniki has been an urban

[1] Environmental Informatics Research Group, School of Mechanical Engineering, Aristotle University, Thessaloniki, Greece and Center for Interdisciplinary Research and Innovation (CIRI-AUTH), Balkan Center, Buildings A & B, Thessaloniki, 10th km Thessaloniki-Thermi Rd, P.O. Box 8318, GR 57001, tkassand@physics.auth.gr
[2] Environmental Informatics Research Group, School of Mechanical Engineering, Aristotle University, Thessaloniki, Greece and Center for Interdisciplinary Research and Innovation (CIRI-AUTH), Balkan Center, Buildings A & B, Thessaloniki, 10th km Thessaloniki-Thermi Rd, P.O. Box 8318, GR 57001, kkara@auth.gr

center and the capital of the Macedonia region for more than 2,500 years, hosting a variety of industrial, transportation and commercial activities. In the second half of the 20th century, the city has expanded, becoming a densely populated area with a low percentage of green spaces.

The KASTOM project

The KASTOM air quality monitoring and forecasting system will be able to provide both primary information (concentrations of regulatory pollutants) and air quality indicators with high spatial accuracy via electronic information services. An innovative emissions modelling method allowing for the estimation of anthropogenic and natural emissions, enriched with a real time emission module to cope for residential heating contributions to the overall urban emissions of the Thessaloniki area. Three-level nesting for meteorological and AQ forecasting has been applied (WRF ver. 5.3 and CAMx ver 6.2 for meteorological and AQ related modelling respectively). The AQ fusion is being developed taking into account modeling results, LCAQSN Network measurements, land use data, estimations of traffic emissions, as well as additional data sources.

LCAQSN Network

The KASTOM project (www.air4me.eu) is currently installing 33 low cost air quality sensor nodes (LCAQSN) in GTA with emphasis in the inhabited areas. Particle sensors for PM10 and PM2.5 are included in all aforementioned nodes (manufacturer: Beijing Plantower Co., Ltd) while 19 nodes include also sensors for gas pollutants (more specifically NO2, O3 and CO, manufacturer: Alphasense Ltd., U.K.), all AQ sensors are uncalibrated. Meteorological parameters (Air Temperature, Relative Humidity and Air Pressure) are monitored with the aid of the BME280 integrated environmental sensor (manufacturer: Bosch Sensortec, Germany). The communication module is based on LoRaWAN and is supported by five gateways.

The quality of the breathed air is monitored in the GTA via the monitoring stations of the official network that belongs to the Ministry of the Environment and is operated by the Prefecture of Central Macedonia, as well as via the networks of the Municipality of Thessaloniki and of the Delta Municipality. We have selected seven stations of the official network to collocate with the KASTOM Nodes as shown in Figure 1. For the scope of the current study, data from one of the collocated nodes were used (Kordelio area).

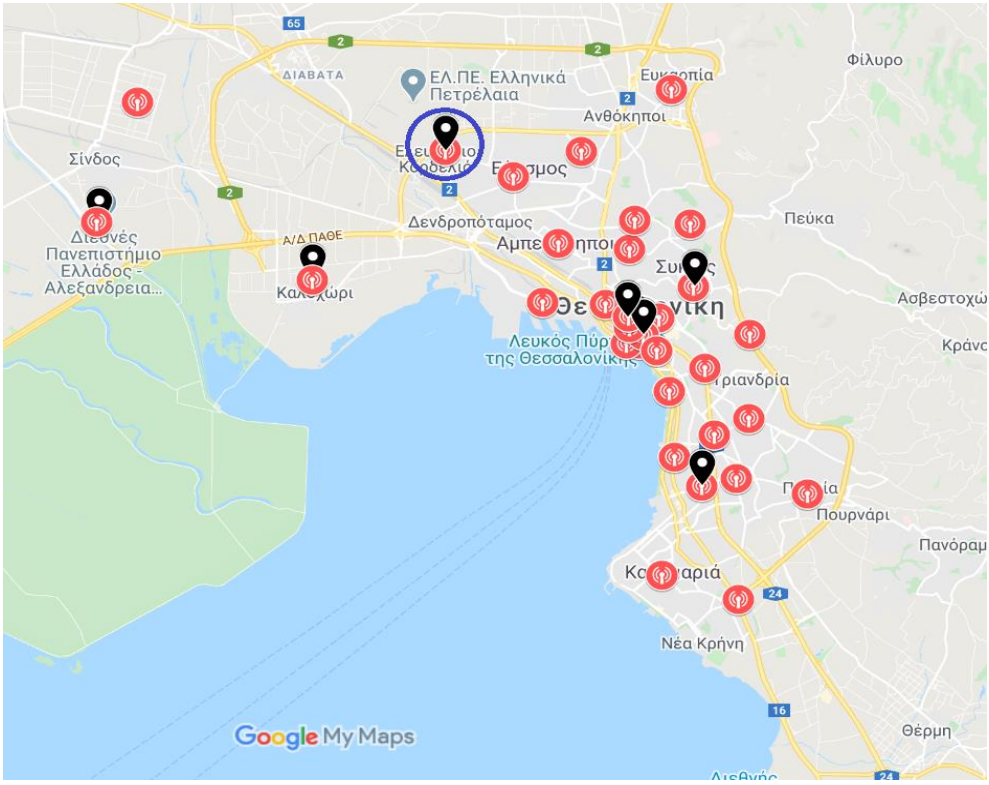


Fig. 1. Map of the Thessaloniki area including 7 reference air quality monitoring stations (black marks) as well as the LCAQSN locations (red marks), left. Within the blue circle are the collocated node and reference station used in this study (background: Google Maps ®).

Data

The initial dataset consists of hourly measurements of five variables of the KASTOM node and the PM10 measurements of the official air quality station (used as the target variable in the modelling procedure) of Kordelio, for an eight-month period between 06/11/19 and 02/07/20, as presented in Table 1.

Node Variables	**Target Variable**
PM10, PM2.5, Temperature, Relative Humidity, Pressure	PM10 (reference)

Table 1. Initial Dataset.

Tools

All computations conducted in WEKA [4] and in Python with the use of Scikit Learn library [5].

2.2. Methods

Feature Engineering

The first step of the computational procedure aimed at generating a set of features, capturing the maximum amount of information, by using the initial dataset. for this reason and for every one of the five node variables (Table 1) the following features where engineered: 24 hour lags, two temporal features under cyclical sine transformation (hour of day, day of week) and rolling mean, minimum, maximum, median, standard deviation, range (maximum minus minimum), minimum/maximum and ratio (minimum/maximum) of the last 6, 12 and 24 hours. On this basis the final dataset consists of 232 Features as presented in Table 2.

Lag	Mean	Min	Max	Median	Std	Range	Ratio	Temporal	Total Number of features
120	15	15	15	15	15	15	15	2	232

Table 2. Number of features in the final dataset.

Feature Selection

Many features introduce noise in data-driven models and can lead to reduced model performance, while simplifying the modeling procedure is better in terms of interpreting the model results and reducing computational time.

A common practice in data driven modeling is feature reduction, using feature selection algorithms. Among such algorithms, wrapper feature selection uses a search method to scan feature space, create different subsets of features and evaluating them using a ML algorithm. The subset of the feature space that minimizes a selected metric, in our case Root Mean Squared Error (RMSE), is selected as the optimal. We therefore applied three wrapper feature selection algorithms resulting in three different feature subsets (Table 3).

Search Method	ML Algorithm	Abbreviation	Number of features selected
Genetic Algorithm	Random Forests	GA subset	80
Particle Swarm Optimization	k - Nearest Neighbors	PSO subset	93
Greedy Stepwise	Regression Trees	GS subset	14

Table 3. Overview of the Feature Selection methods.

Genetic Algorithm (GA) [6] and Particle Swarm Optimization (PSO) [7] are metaheuristic search methods inspired from Charles Darwin's theory of natural evolution and the concept of swarm intelligence often seen in animal groups, accordingly. GA initialized with 0.6 crossover probability, population size and generations equal to 20 and PSO with 0.01 mutation probability, 0.33 social weight and population size and iterations equal to 20. Greedy Stepwise on the other hand performs a greedy forward search through the space of attribute subsets.

Regression trees, M5P [8, 9] is a reconstruction of Quinlan's M5 algorithm for inducing trees of regression models. M5P combines a conventional decision tree with the possibility of linear regression functions at the nodes. K - Nearest Neighbors (k-NN) and Random Forests (RF) will be described in the next paragraph.

Modeling Algorithms

Modelling focused on “PM10 reference” of the official monitoring station, using the 3 feature subsets described previously as input data. The following three algorithms were used in every subset, therefore leading to nine different model implementations.

i) Random Forests (RF) [10], consist of a population of trees (forest) which are trained and then used as an ensemble, i.e. each tree via weighted voting contributes to the final result. We initialized the algorithm with 250 trees.

ii) k-NN [11], the principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point and model its value (label) from these. We initialized the algorithm with 3 neighbors.

iii) A Multi-Layer Perceptron (MLP), is a feedforward artificial neural network that maps sets of input data onto a set of appropriate outputs, consisting of at least 3 hidden layers. We used Adam optimization algorithm [12] and the rectified linear unit activation function.

Ensemble Method

Ensemble modelling combines the predictions of several base estimators to improve robustness and generalizability over a single predictor [13]. In this study, a stacked ensemble procedure with two levels is used. The highest-level ensemble model (level 2), generates the final prediction as the average value, of three high level (level 1) meta-models trained on out-of-fold predictions of the nine base models (level 0), of the overall ensemble model.

The three level 1 meta-models used where RF, MLP and Linear Regression (LR). The final procedure is presented in Figure 2.

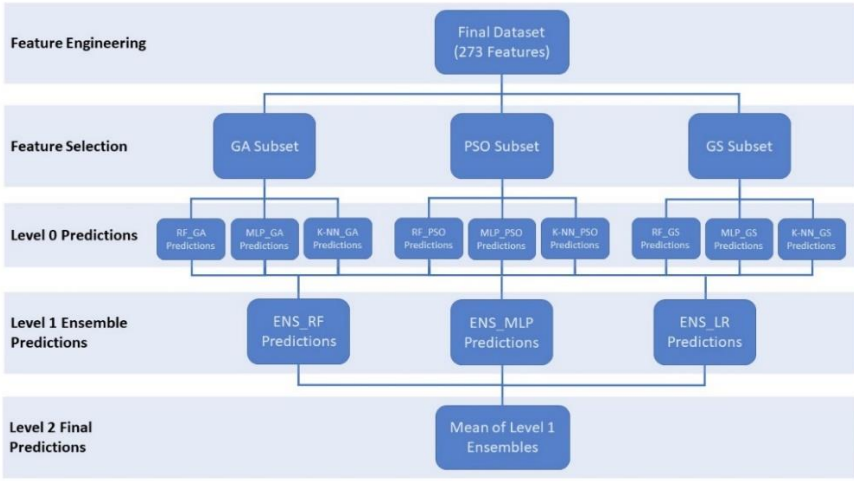


Fig. 2. Overview of the modeling pipeline.

Validation

We evaluated the models using the 10-folds cross-validation method [14]. We used coefficient of determination (R2), Pearson correlation coefficient (r), Index of Agreement (IA), RMSE and Mean Absolute Error (MAE) as evaluating metrics (described in the following equations):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \quad (2)$$

$$IA = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sqrt{\sum_{i=1}^{n}(|y_i - \bar{y}| + |\hat{y}_i - \bar{y}|)^2}} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (5)$$

Here n is the number of samples. $\hat{y}_i$ is the predicted/the candidate method (node) value, and y_i is the real observed/reference measurement value. $\bar{\hat{y}}$ refers to the average predicted value and $\bar{y}$ is the average observed data.

On top of these metrics, the Relative Expanded Uncertainty (REU, Ur) of the measurements was calculated. The European Air Quality Directive [15] defines the Data Quality Objective (DQO) that monitoring methods need to comply with to be used as indicative measurement for regulatory purposes. The DQO is a measure of the acceptable uncertainty for indicative measurements. According to the Directive, allowed uncertainties are 50% for PM10. The REU has been calculated following the methodology described in the Guide to the Demonstration of Equivalence of Ambient Air Monitoring Methods [16], on the basis of the following equations:

$$U_r(\hat{y}_i) = \frac{2\left(\frac{RSS}{n-2} - u^2(y_i) + [b_0 + (b_1 - 1)y_i]^2\right)^{\frac{1}{2}}}{\hat{y}_i} \quad (6)$$

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - b_0 - b_1 y_i)^2 \quad (7)$$

where b_0 and b_1 are the slope and intercept of the orthogonal regression, respectively, RSS is the sum of squares of the residuals, and u is the uncertainty of the reference instrument.

3. Results

We present the evaluation metrics of all models (level 0, level 1 and level 2) together with the same metrics for the raw measurements of the node, in Table 4. The node's performance is very poor especially in terms of errors, while the r values are acceptable, meaning that the node is capable of describing the basic variability of the monitored parameters time, but it overestimates their values, as shown in Figure 3. All the models are increasing the performance in acceptable levels, with every one of them achieving a Ur lower than 50%, while the ensembles models (level 1 and level 2) achieve the best results.

	R^2	**Pearson**	**RMS**	**MAE**	**IA**	**U_r (%)**
RF_GA	0.803	0.92	9.362	5.642	0.955	8.79
RF_PSO	0.79	0.915	9.63	5.846	0.952	9.03
RF_GS	0.8	0.918	9.464	5.728	0.954	8.9
MLP_GA	0.69	0.872	11.65	7.896	0.928	10.9
MLP_PSO	0.689	0.862	12.116	8.178	0.925	11.52
MLP_GS	0.643	0.857	12.266	8.084	0.919	11.38
K-NN_GA	0.698	0.874	11.586	7.022	0.93	10.87

K-NN_PSO	0.658	0.86	12.18	7.455	0.921	11.36
K-NN_GS	0.646	0.85	12.608	7.664	0.916	11.85
ENS_RF	0.821	0.918	9.478	5.7	0.956	9.15
ENS_MLP	**0.834**	0.924	9.158	5.728	0.959	**8.55**
ENS_LR	0.832	**0.925**	**9.045**	**5.52**	**0.96**	8.67
MEAN_OF_ENS	0.833	0.924	9.095	5.551	**0.96**	8.75
Raw Node	-0.286	0.708	105.45	72.713	0.39	133.10

Table 4. Evaluation of all the models and the raw values of node.

We choose to compare the level 2 mean of ensembles predictions and the raw node measurements with the reference measurements in Figure 3, Figure 4 and Figure 5. The final model slightly underestimates the actual concentration levels mainly by being unable to handle extreme values as shown in Figure 3 and Figure 4. The main outcome of this study is that, as shown in Figure 5, the final model is able to "upgrade" the sensor in terms of quality, as it is described by The European Air Quality Directive, yielding an uncertainty lower than 20% in all the cases, while the raw measures are far from the 50% limit.

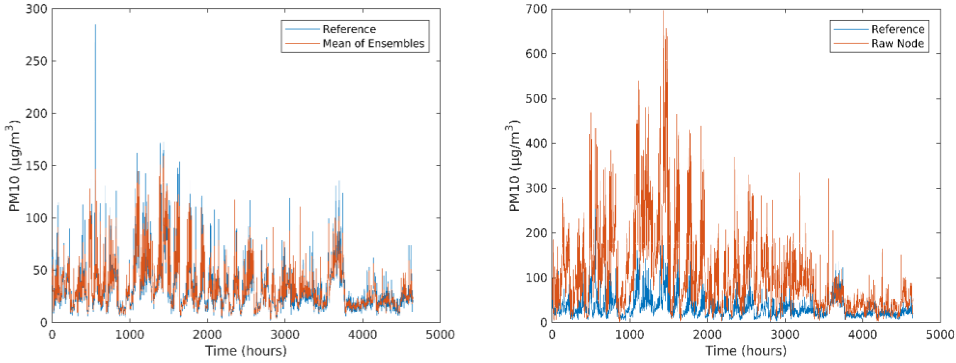


Fig. 3. Reference and Mean of Ensemble Values (left) and Row Node Values (right) through time, for the study period.

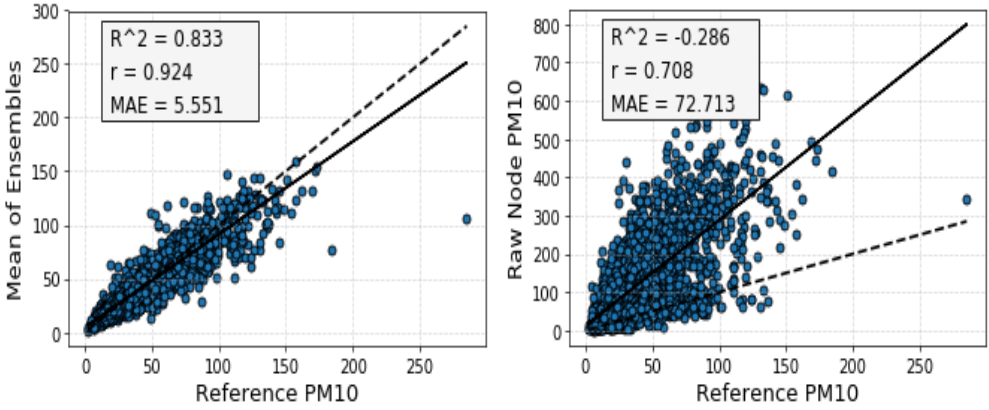


Fig. 4. Scatter plots of Reference vs Mean of Ensembles (left) and Raw Node Values (right). The dotted line corresponds to the ideal performance and the straight line to the actual.

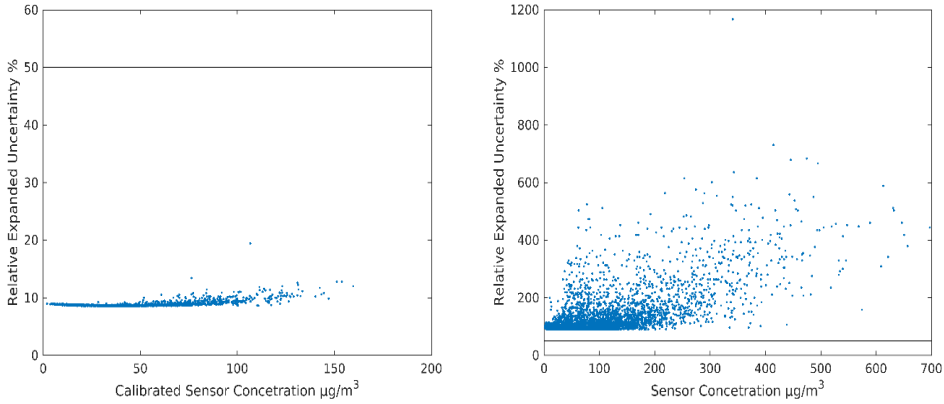


Fig. 5. Relative Expanded Uncertainty of Calibrated Sensor Values (left) and Raw Values (right). The horizontal line is the 50% limit, under which measurements considered indicative.

4. Discussion

The modelling methodology (pipeline) presented and tested in one sensor of the KASTOM LCAQSN network is yielding positive results and is able to upgrade the node in terms of accuracy and uncertainty, and therefore classify its measurements as indicative, which is currently the official category for LCAQSNs trustability. The level 2 ensemble model doesn't lead to better results of the level 1 ensembles, and therefore this step can be skipped in the future, on the basis of additional computational experiments to be conducted. Moreover, the duration of the study period is very short, and more measurements (ongoing) are necessary for achieving better model training results.

The initial dataset contains only 5 variables, and although feature construction expanded the feature space, more variables, that are operationally available should be included, such as meteorology forecasts, land use and traffic indicators.

Lastly, and in order to improve a sensor's performance a generalized model should be built, taking account all collocated nodes and trained in more data, while tested both on time and space, leaving out time periods and stations from the training phase. Results of this study indicate that such a model could be feasible for the GTA. The performance of such model will be determined by the representativeness of the measurements and different models should be considered for groups of sensors with similar behavior.

Acknowledgments

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE. Project code T1EDK-01697; project name Innovative system for air quality monitoring and forecasting (KASTOM, www.air4me.eu).

References

[1] Spinelle, L., Gerboles, M., Villani, M.G., Aleixandre, M., Bonavitacola, F., (2015): Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: ozone and nitrogen dioxide. Sensor. Actuator. B Chem. 215, 249–257.

[2] Spinelle, L., Gerboles, M., Villani, M.G., Aleixandre, M., Bonavitacola, F., (2017): Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO 2. Sensor. Actuator. B Chem. 238, 706–715.

[3] De Vito, S., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R., Di Francia, G., (2018): Calibrating chemical multisensory devices for real world applications: an in-depth comparison of quantitative Machine Learning approaches. Sensor. Actuator. B Chem. 255, 1191–1210.

[4] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016): The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[5] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[6] David E. Goldberg. (1989): Genetic Algorithms in Search, Optimization and Machine Learning (1st. ed.). Addison-Wesley Longman Publishing Co., Inc., USA.

[7] Moraglio A., Di Chio C., Poli R. (2007): Geometric Particle Swarm Optimisation. In: Ebner M., O'Neill M., Ekárt A., Vanneschi L., Esparcia-Alcázar A.I. (eds) Genetic Programming. EuroGP 2007. Lecture Notes in Computer Science, vol 4445. Springer, Berlin, Heidelberg

[8] Ross J. Quinlan: Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, 343-348, 1992.

[9] Y. Wang, I. H. Witten: Induction of model trees for predicting continuous classes. In: Poster papers of the 9th European Conference on Machine Learning, 1997.

[10] Breiman, L. (2001): Random Forests. Machine Learning 45, 5–32.

[11] Cover, T.M., & Hart, P.E. (1967): Nearest neighbor pattern classification. IEEE Trans. Inf. Theory, 13, 21-27.

[12] Kingma, Diederik & Ba, Jimmy. (2014): Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.

[13] Wolpert, David. (1992): Stacked Generalization. Neural Networks. 5. 241-259. 10.1016/S0893-6080(05)80023-1.

[14] Stone, M. (1974): Cross-Validatory Choice and Assessment of Statistical Prediction (with Discussion). Journal of the Royal Statistical Society (Series B), 36, 111-147.

[15] EU, 2008. Directive 2008/50/EC of the European Parliament and the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe and Directive (EU) 2015/1480 of 28 August 2015 on Reference Methods, Data Validation and Location of Sampling Points for the Assessment of Ambient Air Quality.

[16] EC WG, 2010. Guide to the Demonstration of Equivalence of Ambient Air Monitoring Methods. Report by EC Working Group on Guidance. Available at: ec.europa.eu/environment/air/quality/legislation/pdf/equivalence.pdf.

Online energy forecasts for the Internet of Things

Danja Steinberg[1], Julien Murach[2], Achim Guldner[3], Klaus-Uwe Gollmer[4]

1. Introduction

Only a few years ago, the main focus of society was on using the cheapest possible electricity. With the changes in green thinking, many have redirected their focus away from saving money to saving resources and thus primarily to protecting the environment. However, questions arise, like "how can you find out what the energy is currently composed of" or "what even is a good and a bad CO_2 / kWh emission value"? Is it possible to make predictions like: "Will it be possible to use low-CO_2 energy in the next few hours?"

Once this information has been determined, the next step is to react to it as automatically as possible, e.g. to switch on an electrical consumer, if green energy is available in sufficient quantity. The possible application scenarios here are very diverse and often depend on the personal preferences of the individual user. Our idea, the integration into an existing IoT platform for training and rapid prototyping in an industrial environment, enables the creative implementation of own ideas. The IoT^2-workshop links the Internet of Things (IoT) and Thinking (EDGE-AI) with a graphical programming language and thus enables people with little to no knowledge in computer science to use the possibilities of distributed systems.

The proposed tool provides the percentage of electricity produced by renewable energy sources and the current CO_2 values per energy unit. A forecast for the next 24 hours is also available. The information is updated every 15 minutes and allows for a general idea on how to organize the daily life, e.g. when to turn on a washing machine or charge your phone to live as environmentally conscious as possible.

Ideally, this information would be made available online, e.g. via an API (Application Programming Interface) from the local network operator. Various platforms already offer forecast data for Germany, but unfortunately not in the desired level of detail. The work is therefore divided into the following steps:

1. Querying of the available energy data (SMARD) and compilation of the information needed to calculate the carbon footprint of the individual energy sources.
2. Making the results easily available on an open-source platform, accessible online for the use with Internet of Things Devices (ThingSpeak)
3. Integration into the IoT^2-workshop and implementation of exemplary usage scenarios

To carry out the first and second step, a Raspberry Pi 3 Model B+ is programmed using Python and CRONTAB.

[1] Environmental Campus Birkenfeld, Germany, s18d85@umwelt-campus.de
[2] Environmental Campus Birkenfeld, Germany, s19ade@umwelt-campus.de
[3] Environmental Campus Birkenfeld, Germany, a.guldner@umwelt-campus.de
[4] Environmental Campus Birkenfeld, Germany, k.gollmer@umwelt-campus.de

ThingSpeak is hosted at the computer center of the university. The application is implemented in the form of a microcontroller (IoT-Octopus). Figure 1 illustrates the interaction of the distributed system.

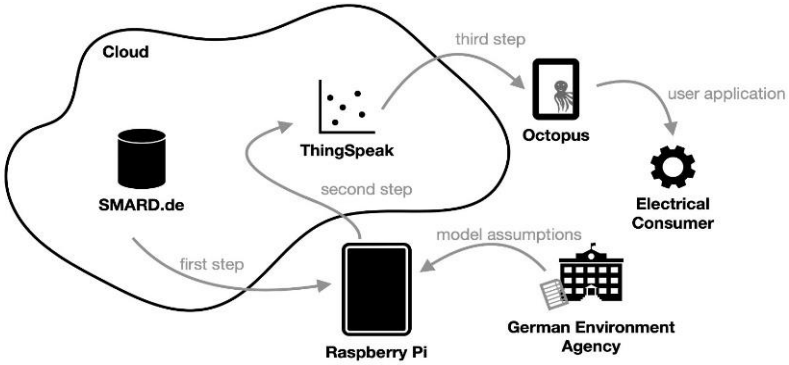


Fig. 1. Overview of all components and steps within the project.

It should be noted that the third step and user application is made simple and flexible, by using a visual programming language (ArduBlock)[5]. This also allows the easy integration of additional sensors and actuators to control e.g. electrical consumers. A visual programming language is characterized by the fact that a program, algorithm, or system behavior is represented as a graphical element or block and the overall function is defined by the arrangement of these blocks. Associated programming examples for getting started are available on the website of the IoT²-workshop[6]. The open source IoT²-workshop is free of charge and also provides help for a quick and easy start, more project ideas and specially designed blocks. Among them is also a block to use the function described in this paper [1][2].

Similar approaches

There are already a few websites that offer similar data. An example is the website "electricityMap"[7]. According to their webpage, they are "the world's most comprehensive platform measuring the carbon footprint of your electricity consumption" and offer a worldwide coverage of high frequent comprehensive data. On the page you can learn a lot about the current state of energy production, the currently produced CO_2, and the percentage of renewable energy. However, it is only possible to see the current data and the last 24 hours. Furthermore, the data can only be downloaded after registration and paying a fee or using a student account. It is a very interesting website to learn more about where the energy is currently coming from and how regenerative it is. Nevertheless, the predicted data is not freely accessible for everyone and thus people cannot easily use it to plan their everyday life.

[5] http://blog.ardublock.com/ (28.07.2020)
[6] https://iot-werkstatt.de (28.07.2020)
[7] https://api.electricitymap.org/?utm_source=electricitymap.org&utm_medium=referral#about (28.07.2020)

2. Methods

In the following, all sources used for the creation of this tool are explained and evaluated. Furthermore, we also explain how the results are displayed and how they can be processed or used to help people become more environmentally friendly.

2.1. Source of the data SMARD

Due to the EU commission regulation No 543/2013 from June 14th, 2013 on the transmission and publication of data on electricity markets [3], the German government aims for more transparency regarding the energy production. At the moment, there are a few websites where you can get the current information about the energy production. The most popular websites are SMARD.de [4], Energy charts from Fraunhofer ISE [5] and entso-e [6]. The website SMARD.de will be used in this project to obtain the needed data on electricity generation. This has the advantage that the data is provided directly by the federal network agency and therefore has a high credibility and accuracy. Furthermore, the data is freely accessible to everyone and can be downloaded without prior registration. SMARD stands for "Strommarktdaten" (electricity marked data). It aims to increase the transparency of electricity generation and provides a comprehensive overview of what is happening in the electricity market. The electricity info graphic regarding Germany and parts of Europe can be displayed here almost in real-time [4].

Data can be obtained regarding the categories "power generation", "power consumption", "market" and "system stability". In the "power generation" category, the system is further subdivided into current- and forecast power generation. A real-time API is still missing, but the data is available in computer-readable file formats (XML, XLS, or CSV) [4].

The data of the currently generated energy is divided into 12 different energy sources, which can be seen in the downloaded CSV file in Table 1.

Date	Time of day	Biomass[MWh]	Hydropower[MWh]	Wind offshore[MWh]	Wind onshore[MWh]	Photovoltaics[MWh]	Other renewable[MWh]	Nuclear[MWh]	Fossil brown coal[MWh]	Fossil hard coal[MWh]	Fossil gas[MWh]	Hydro pumped storage[MWh]	Other conventional[MWh]
Jul 9, 2020	4:00 AM	1,098	582	75	580	0	34	1,334	2,665	587	2,298	40	193
Jul 9, 2020	4:15 AM	1,095	589	78	614	0	34	1,333	2,664	588	2,315	38	193
Jul 9, 2020	4:30 AM	1,094	592	78	663	0	34	1,333	2,669	617	2,319	38	193
Jul 9, 2020	4:45 AM	1,09	611	79	712	1	34	1,334	2,696	644	2,321	51	193
Jul 9, 2020	5:00 AM	1,091	578	78	771	4	34	1,334	2,725	706	2,311	35	193
Jul 9, 2020	5:15 AM	1,096	578	79	832	13	34	1,334	2,78	725	2,314	27	193
Jul 9, 2020	5:30 AM	1,096	572	81	892	31	34	1,334	2,806	749	2,297	26	192
Jul 9, 2020	5:45 AM	1,096	584	85	942	71	34	1,334	2,836	778	2,302	34	192
Jul 9, 2020	6:00 AM	1,099	567	88	999	144	34	1,333	2,809	900	2,294	46	192
Jul 9, 2020	6:15 AM	1,099	595	87	1,057	244	34	1,334	2,824	975	2,282	48	192

Table 1. Current electricity generation distributed among the individual energy sources (source: smard.de).

In order to determine the share of renewable energies, a classification of the different energy sources must be carried out. We have classified “biomass”, “hydropower”, “wind offshore”, “wind onshore”, “photovoltaics”, “hydro pumped storage”, and “other renewable” as renewable and “nuclear”, “fossil brown coal”, “fossil hard coal”, “fossil gas”, and “other conventional” as conventional.

Date	Time of day	Total[MWh]	Wind offshore[MWh]	Wind onshore[MWh]	Photovoltaics[MWh]	Other[MWh]
Jul 10, 2020	4:00 AM	38,187	205	1,409	0	31,403
Jul 10, 2020	4:15 AM	-	211	1,458	0	-
Jul 10, 2020	4:30 AM	-	217	1,507	0	-
Jul 10, 2020	4:45 AM	-	222	1,556	0	-
Jul 10, 2020	5:00 AM	40,333	230	1,608	4	32,534
Jul 10, 2020	5:15 AM	-	235	1,641	22	-
Jul 10, 2020	5:30 AM	-	241	1,675	66	-
Jul 10, 2020	5:45 AM	-	246	1,705	127	-
Jul 10, 2020	6:00 AM	46,912	251	1,738	206	37,328
Jul 10, 2020	6:15 AM	-	259	1,752	303	-

Table 2. Forecast electricity data in total and from the energy sources “wind offshore”, “wind onshore”, and “photovoltaic”.

For a prediction of the selected CO_2 production per kWh and the percentage of renewable energies, a forecast of the data on the generally produced power is necessary. On the SMARD platform, such a forecast is published for the following day (cf. Table 2). This forecast does not include the complete breakdown as in the case of generating energy sources. In the forecast, a distinction is made between forecast total output and energy production from “wind offshore”, “wind onshore”, “photovoltaics”, and “others”, whereby “others” consist of the sum of the remaining energy production types.

2.2. Step 1: Raspberry Pi setup and data processing

Because the SMARD platform does not currently have an API, we use the Selenium library[8] for the download of the required data.

The total of all renewable energies is set in relation to the total energy production. Thus, the value for the current percentage of renewable energy is created. To be able to make a forecast for the next 24 hours, the forecast data is downloaded. Since not all renewable energies are listed here, only the forecast values for “wind offshore”, “wind onshore”, and “photovoltaic” are summarized. The remaining renewable energy source values are taken from the previous file, as no clear predictions can be made about it. The sum can now again be offset against the forecast total energy. This is done for all values of the next 24 hours to create a good overview of the day.

[8] https://github.com/SeleniumHQ/selenium (28.07.2020)

For the calculation of the CO_2 emission every 15 minutes, it must first be determined how much gCO_2/kWh is produced, depending on the type of production. Therefore, current conversion figures are required. We have taken these values from the current publication of the "Emissionsbilanz erneuerbarer Energieträger" (Emission balance of renewable energy sources) [7]. Table 3 shows the relevant values for the used energy sources.

Constant values 2018	
Energy source	CO2 [g/kWh]
Biomass	153,605
Hydropower	2,702
Wind Offshore	5,998
Wind Onshore	10,497
Photovoltaics	66,730
Other renewable	50,974
Nuclear	22,370
Fossil brown coal	415,190
Fossil hard coal	390,370
Natural gas	242,650
Pump storage	25,064
Other conventional	315,470

Table 3. CO_2 conversion rates of the different energy sources needed for the calculation of the current CO_2 emissions.

With these conversion rates, it is possible to calculate the expected CO_2 emission. For this purpose, the previously calculated values of the energy sources are multiplied by their CO_2 [g/kWh] conversion rates. Finally, the script calculates the sum, so that the total CO_2 emission is determined. By means of the total energy values for this 15 minutes, one can now determine how much CO_2 is emitted for the energy production.

In order to calculate the predictions, the current value for the projected energy sources ("wind offshore", "wind onshore", and "photovoltaic") is recalculated and the old values for the other energy sources are used.

2.3. Step 2: Data upload

The processed and calculated data can now be displayed in a user-friendly way. It needs to be accessible for everyone and easily understandable. Furthermore, the platform must have a universal real-time API, so that the data upload can be integrated into the IoT^2-workshop.

To achieve this, we use ThingSpeak, an open-source Internet-of-Things application, which was originally launched by ioBridge in 2010. The data can be visualized and further processed by users and exported in a computer-readable way [8]. This makes it possible to, for example, send an e-mail when the value hits a certain threshold. Further advantages using ThingSpeak arise from the possibility to create a channel, which is either private or public, depending on the data and its privacy policy [9].

Two public channels were created, in order to present the evaluated energy data. The first channel displays several figures for the percentage of renewable energy for now and for the next 7 hours. The second channel supplies figures with the gCO_2 per kWh for the same period of time.

2.4. Step 3: IoT2-workshop and application

Another important point is the direct application. Therefore, we implemented a new block, which can access our ThingSpeak channels. The microcontroller of choice is the "IoT-Octopus"[9]. The Octopus was also co-developed at the Environmental Campus Birkenfeld. It is based upon an ESP8266, a low-cost Wi-Fi microchip. The Octopus provides practical interfaces for sensors and actuators and makes using them even easier [1][2]. We are using a Charlie-Wing LED matrix[10] to present our results visually and directly in an understandable way. The matrix can be easily connected to the top of the Octopus and can either be used to display a trend graphic or show a certain text, e.g. the current CO_2 emissions from the energy production. It is now possible to place the Octopus anywhere in the house and use it as guidance for the electricity and its sustainability.

3. Results

We have implemented the first functions that make it possible to see the current percentage of renewable energy[11] and CO_2 emissions on the website [12]. Furthermore, the prediction of both values in the span of the next 7 hours can be seen. With the help of these flow charts, it is now possible to easily base certain decisions on the current energy mix and resulting CO_2 emissions. For example, the washing machine might be turned on one or two hours later when more renewable energy is available.

Figures 2 and 3 show the data aggregated on July 21st and the current and predicted values from that day. The y – axis of the figures contains the percentage of the renewable energy/the CO_2 emission in gCO_2/kWh and the x – axis contains the corresponding date and time.

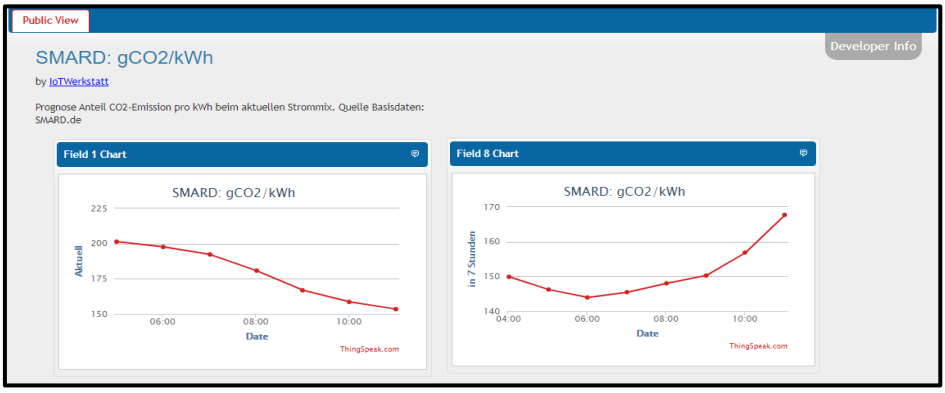


Fig. 2. Screenshot from the ThingSpeak website, showing the public channel "Percentage Renewable Energy". This channel aggregates the current percentage of renewable energy and a forecast for the next 7 hours.

[9] https://www.tindie.com/products/FabLab/iot-octopus-badge-for-iot-evaluation/ (28.07.2020)
[10] https://learn.adafruit.com/adafruit-15x7-7x15-charlieplex-led-matrix-charliewing-featherwing (28.07.2020)
[11] http://thingspeak.umwelt-campus.de/channels/688 (28.07.2020)
[12] http://thingspeak.umwelt-campus.de/channels/689 (28.07.2020)

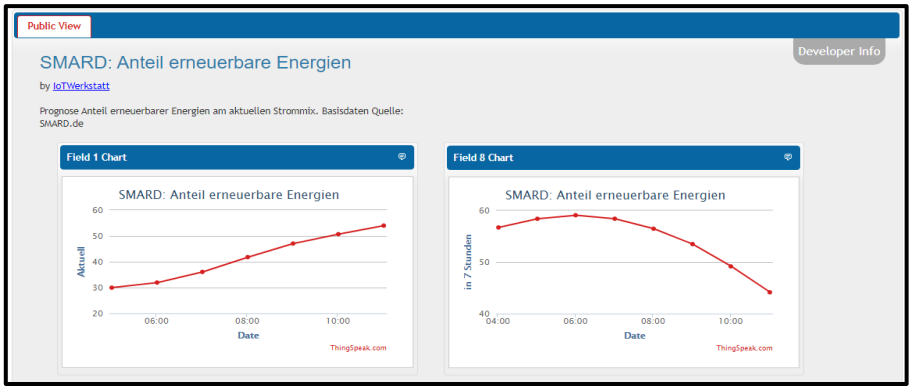


Fig. 3. The public channel "CO_2 Emission". This channel aggregates the current CO_2 emission and the forecast for the next 7 hours.

It is now possible to further process the shown data in ThingSpeak. Using the ArduBlock example from Figure 4, users can immediately see when the power generation produces the least amount of CO2 for the next 7 hours, allowing them to schedule their activities accordingly. In this case, the "greenest" energy will be available in 1 hour as can be seen on the display.

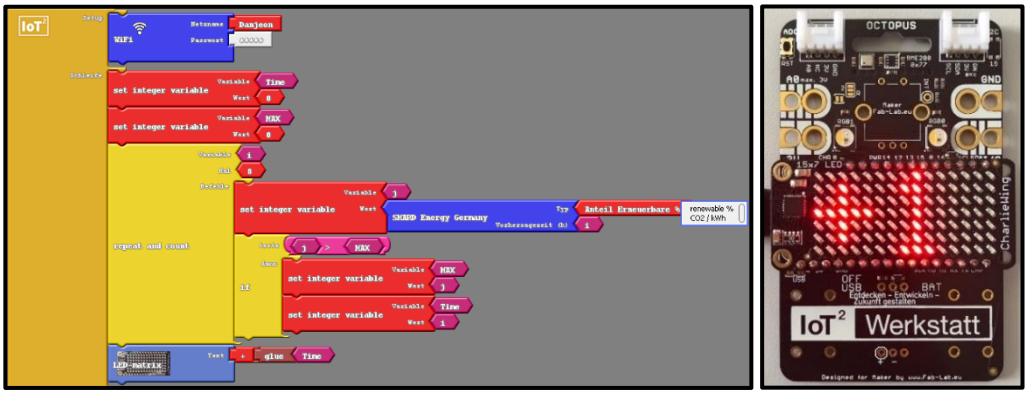


Fig. 4. (Left) The implemented ArduBlock "SMARD Energy Germany" is used to display the uploaded values from ThingSpeak visually on a LED matrix on an Otcopus (Right). The Otcopus with the LED Matrix is displaying the time with the highest percentage of renewable energy. This allows users to instantly see when the highest percentage of renewable energy will be available.

The program can be extended to also display the according percentage of renewable energy. This can be used to present the information in an educational and informative way. For example, we built a prototype to display the octopus and a short explanation, to exhibit it around the university.

Fig. 5. An educational prototype with the tool and an explanation of its relevancy. The Otcopus with the LED Matrix is displaying the time with the highest percentage of renewable energy and the according percentage of the renewable energy alternately.

With the ArduBlock, it is possible to program many different applications, for example start a motor or other electrical device when the highest percentage of renewable energy is available.

4. Discussion

So far, we are able to evaluate and upload the data according to our selected criteria. The created tool is operational and can help people live a more ecological life. It can be concluded that the most environmentally friendly time to start the washing machine on the 21st of July would have been at around 1 pm (which can be seen in Figures 2 and 3 forecast for “in 3 hours”). Here, the percentage of renewable energy is at local maximum at about 56 % and the CO_2 emission is at a local minimum at about 145 gCO_2/kWh.

In general, it is very easy to derive the interaction of the renewable energies and the CO_2 emission from the graphic designer. In the morning hours, the percentage of renewable energies is still quite low and therefore the CO_2 emission is high. Around 5 to 6 am, when the sun rises, the percentage of renewable energies increases and the emission decreases. From about 15 o'clock, the percentage of renewable energies decreases, and the emission rate increases again. Furthermore, we are able to access the data using the ArduBlock and further process it, or rather, use it to control processes.

However, there are still some things that have to be taken into consideration using the tool or could be added in the future. The first problem that will be discussed and revised in the future is the accuracy of calculated values, as they are rather rough. We determine much more a mean than an exact value. In the future, a more precise evaluation of each individual power plant should be aimed for. Thus, the respective raw data and the characteristics of the SMARD site must be taken into account and incorporated into the result.

One little add-on for the ThingSpeak website is the visualization through a kind of traffic light. The light would be green if there is more renewable energy being used or if a certain percentage of renewable energy is predominant. This makes it even easier for users to interpret the displayed percentage without having to be accustomed with the topic. We have looked at the values over a certain period of time and would set this threshold at about 60 %.

It can be seen that there still is room for further optimisation, but the tool may be a next step when it comes to helping people become more environmentally aware.

Future potential

With the growing trend of the digital transformation and the increasing necessity of environmental protection, there are more and more technologies for a conscious and more sustainable everyday life. One example of this trend is the smart home, which has become more and more popular in recent years. This follows the basic idea that you are saving energy and live environmentally conscious.

But if we want to use these new kinds of technologies like our tool, society must first adapt to this new way of life. Especially ordinary habits will have to be adjusted to this new situation. An easy example of this needed mental shift is doing laundry. Instead of doing this task at a certain time to ensure that it is finished at the desired time, the washing machine would be set up in the morning. The laundry would then be washed when it is most energy efficient. So, the general idea that everything must be done right now to be effective needs to change. These changes in the daily routine will require some time to get used to but are necessary for a greener future.

References

[1] Gollmer, K.-U.; Burger, G.; Krieger, R. & Wahl, M.: IoT-workshop: Application examples for adult education. Chapter in Big Data Analytics for Cyber-Physical Systems, Machine Learning for the Internet of Things, 1st Edition, Elsevier, Paperback ISBN: 9780128166376

[2] Gollmer, K.-U.; Kreten, S.; Stolz, F.; Dartmann, G. & Burger G.: IoT-Workshop: Blueprint for pupils education in IoT, Chapter in Big Data Analytics for Cyber-Physical Systems, Machine Learning for the Internet of Things, 1st Edition, Elsevier, Paperback ISBN: 9780128166376

[3] EUR-Lex (2013): https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32013R0543 (21.07.2020)

[4] Bundesnetzargentur (2020): https://smard.de/home/Ueber_SMARD/668 (21.07.2020)

[5] Fraunhofer ISE (2020): https://www.energy-charts.de/index_de.htm (27.07.2020)

[6] Entsoe (2020): https://transparency.entsoe.eu/dashboard/show (27.07.2020)

[7] CLIMATE CHANGE (2019): Emissionsbilanz erneuerbarer Energieträger – Bestimmung der verschiedenen Emissionen im Jahr 2018

[8] Marcello M. (2014): ThingSpeak – an API and Web Service for the Internet of Things

[9] GitHub (2020): https://github.com/iobridge/thingspeak (21.07.2020)

Analysis and modeling of low-cost air quality sensor data towards their computational improvement

Petros Zimianitis[1], Kostas Karatzas[2]

1. Introduction

According to the World Health Organization (W.H.O.) 91% of the world's population lives in places where air quality exceeds WHO guideline limits [1]. Hence, drastic actions are needed to improve air quality worldwide. In order to do that, we need to continuously monitor air pollutants.

Due to that need, air quality (AQ) monitoring systems of high quality are deployed in various areas. Although very accurate, they are not cost effective. The development of IoT technologies and the introduction of low-cost air quality sensors support AQ monitoring, yet accompanied by the compromise concerning the quality measurements. Such quality problems are attributed to the limitations posed by the measuring principles applied by low cost AQ sensors and various operational and environmental influences. The goal of this study is to suggest a data-driven computational method in order to improve those low-cost systems, by using machine learning algorithms. The algorithms employed in this study are used for the development of data-driven models that receive as input the measurements conducted by the low cost AQ systems and produce as output an arithmetic estimation of the concentration of the pollutant(s) of interest (here, PM10 concentrations). The whole procedure (data preparation, feature selection and training) was implemented using Matlab.

2. Materials and Methods

2.1. Area of study

The city of Thessaloniki is situated in Northern Greece and the Southern part of the Balkan area. It is the second largest city in Greece with a population of approximately 1 million citizens and it is very densely populated. Traffic as well as industrial emissions and local human activities are the main sources of air pollutants in the area.

2.2. Air Quality Systems

2.2.1. Low cost air quality systems (LCAQS)

Low cost AQ sensors are using different methods (light scattering, chemical reactions etc.) in order to estimate pollutant concentrations [2]. Such sensors are usually installed within an enclosure, are controlled by a smart unit, and communicate their measurements via a communications module, while in some cases

[1] Aristotle University of Thessaloniki, Greece, zimianitisp@gmail.com
[2] Aristotle University of Thessaloniki, Greece, kkara@auth.gr

also storing measurements locally. All these components are called low cost air quality devices or systems (LCAQS). The LCAQS used for this study are located on top of a University building within the University campus in the center of Thessaloniki. The building is situated by Tritis Septrevriou street (40.62 N, 29.22 E). The building's roof is approximately 20m above the ground (about 50m above sea level). More specifically, the Comonet LCAQS (www.comonet.eu) is used in this paper. It is equipped with a PM3003 sensor by Plantower, for measuring PM concentrations (PM1, PM2.5 and PM10) and with a Sensirion SHT11 sensor for measuring temperature and RH. This system is designed by Comonet as a pilot product and it is not yet available in the market.

2.2.2. High quality AQ monitoring system

The high-quality PM10 concentration measurements originate by a high accuracy measurement station manufactured by Aeroqual (www.aeroqual.com). The station comprises two units: the AQM60 unit for gaseous pollutants and the Dust Sentry PM10 (DS-PM10) unit for particulate matter, which is used in the current intercomparison study. The latter includes PM10 sensing capabilities, based on a light scattering nephelometer that has a sharp cut cyclone and an on-board temperature sensor to correct for thermal drift, sheath air filter to keep the optics clean, automatic baseline drift correction, and a fiber optic span to enable a check of the optical components. The DS-PM10 also includes omni-directional heated sampler inlet, sample pump and volumetric flow controller. The unit is MCERTS certified, and South Coast Air Quality Management District (SCAQMD) Rule 1466 pre-approved.

2.3. Preparation

The dataset used for this research contains hourly measurements (time series data with 4019 timestep registrations) from the above-mentioned system for a period of approximately six months (4/9/2018 – 28/2/2019).

The most common method for training and testing a machine learning model is k-fold cross-validation (k-fold CV). This method is the best for utilizing all the available data efficiently, particularly when there is shortage of them [3]. When k-fold CV is performed, several indexes from the initial set are removed. Those are used to form the test set and the remaining are the training set. If this procedure takes place using a time series dataset, then it is expected that during the modelling test phase there will be a certain information leakage. Therefore, the model will already be exposed to the values that are targeted, and therefore "know" the optimum answer. In order to tackle this problem, we followed a procedure different from the classic k-fold CV: the initial training set contained measurements of all the available parameters that the Comonet LCAQS can provide as well as lagged measurements of the same parameters. A lag order of 12 hours was preferred for this study. This set was split into multiple subsets, in a way that consecutive inputs of those subsets have a time difference of 12 hours, and thus they don't contain any "knowledge" that may hamper the performance assessment procedure of the models.

It is not uncommon, that many machine learning algorithms yield better results when normalized data are used for training. For this research, normalized data trials took place for all the proposed algorithms. However, the results did not show significant performance improvement, partly because the actual values of all the variables are of very close order. Therefore, raw data were used for training.

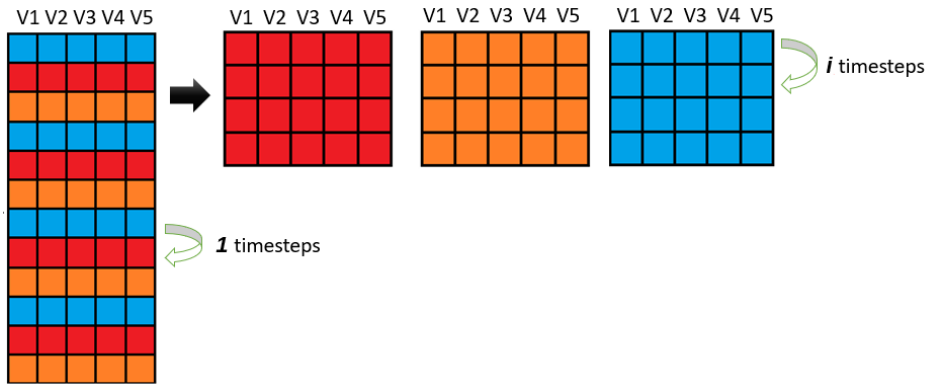


Fig. 1: Representation of the data preparation procedure for training. The initial dataset on the left is split into multiple subsets (red, orange and blue sets). Each subset is used for training of a model. After all models of an algorithm type have been trained, they are used as an ensemble and the PM10 concentration estimation results as the average value of these models.

2.4. Feature Selection

Feature selection is a very common procedure in machine learning and statistical analysis. It is also known as attribute selection or variable selection. This procedure is used to select the most relevant variables or predictors for a model for many reasons of simplicity, running and training time, generalizability etc. The most important reason is to prevent overfitting to the training data or to prevent divergence from the desired output due to irrelevant input variables.

Three feature selection methods were used in this study. The first one is the Spearman correlation coefficient, due to its capability to identify non-linear relationships between the variables [5]. The Spearman coefficient does not assume specific distributions for the variables of interest, although the relationship must be monotonic.

A Genetic algorithm is used as a second method. A Genetic algorithm (GA) is an algorithm inspired by the theory of natural selection. GA are widely used in order to reduce dimensionality in machine learning by operating on a randomly generated initial population of chromosomes (inputs). In an iterative process, the GA produces new generations of chromosomes, after evaluating their fitness (through a fitness function). The last (final generation), contains the chromosomes with the genes that appear to be the strongest. The genes represent the variables of the problem and the strongest are the ones that are selected by the GA [6].

The third method is a bagged-trees algorithm. Although such algorithms are used for predictive modelling, they possess strong feature selection capabilities. During the model training, many algorithm implementations can rank the input variables based on the information they provide to the model in order for it to produce the desired output. We can therefore extract that feature importance ranking and use this information for feature selection [7].

To make the whole procedure as automated as possible, the above-mentioned feature selection methods are used as an ensemble. Each method selects a subset of the predictor variables. Those that are selected by at least 2 of the methods (GA, Bagged-trees, Spearman correlation coefficient) are used to formulate the subset of the variables that is used as an input for the models.

2.5. Machine learning algorithms

Three machine learning algorithms were used in this study: Feedforward neural networks (FFNN), Random Forest algorithms (RF) and Long Short-term Memory neural networks (LSTM NN).

Artificial Neural Networks (ANNs) are machine learning algorithms inspired by the biological neural networks that constitute our brain and the FFNN is its simplest form [8]. Their artificial neurons require inputs and target outputs to be trained and make predictions. FFNNs operate by weighing their inputs according to the connection's weights $\boldsymbol{w}$ (all neuron connections are weighted relative to their importance). The sum of the weighted values is then fed into the propagation (here a sigmoid $\boldsymbol{\sigma(x)}$) function to produce the output (taking into account a bias $\boldsymbol{b}$). During training, the weights and the biases are calculated and readjusted by minimizing the error between the predicted value by the network and the actual value (target). The ANN contains 3 layers: an input layer, a hidden layer of 10 neurons and an output layer

Random forests are an ensemble learning method that is used for classification and regression problems [7] [9]. The models using this kind of method consist of multiple decision trees. During training their goal is to develop a series of rules for the input variables leading to the prediction of a variable of interest. Each tree selects split or cut points for every variable that describes the problem, based on a cost function. Minimizing this function (usually the sum of errors for regression problems) yields the best cutoff points for the algorithm. The final prediction is the result of a voting procedure, in that, the prediction that collects the majority of the votes of the trees the form the forest is the final prediction of the algorithm. For this study, the RF contain 200 trees.

LSTM NNs are recurrent neural networks (RNNs) that are widely used in speech, handwriting and image recognition. Their architecture is unique because they can "remember" past outputs and thus, they have a great advantage when it comes to learning from time series. LSTM neurons contain 2 more functionalities compared to typical neurons, the input gate and the forget gate [10]. They are both responsible for the memory of the neuron as in how much information is stored and how much needs to be forgotten or discarded. The LSTM models that were produced contain 7 layers. An input layer, 2 LSTM layers with 800 neurons, a dropout layer, a fully connected layer with the size being the amount of the selected features, a fully connected layer with 1 neuron and an output layer, following the suggested algorithm setup in Matlab [14].

2.6. Performance Assessment Methods

The models resulting from the use of the aforementioned algorithms are assessed for their performance by their Mean Absolute Error (MAE), Coefficient of determination (R^2) and their uncertainty of the air quality parameters resulting from the model application.

The Mean Absolute Error is a statistic for measuring the error between two samples of the same type. This is a very useful test for models due to the result being of the same order as the model's predictions.

The coefficient of determination is a very common way to assess the performance of machine learning algorithms in regression problems. It is a measure that depicts how well the independent variables can explain the variance of the predicted values. The R^2 value has a range between 0 and 1. A value close to 1 is considered to be very good for the model's performance [11].

The relative (expanded) uncertainty of the measurements or predictions is one of the basic criteria for accepting and characterizing a measured value in air quality monitoring [12]. In order to better assess the performance of every model the relative expanded uncertainty was calculated using orthogonal regression of the predicted values of the models with respect to the observed ones. The calculated uncertainty is then compared to the 50% limit introduced for PM10 measurements by the European Commission [13]. This is very important because it means that the LCAQS, after its computational improvement, can be used as a complementary method for PM10 monitoring.

3. Results

After applying the feature selection techniques, the 21 predictor variables that were selected for the LCAQS were the following:

T, PM2.5(-1h), RH(-1h), PM10(-2h), PM2.5(-2h), PM1(-2h), RH(-2h), T(-2h), PM2.5(-4h), T(-4h), PM10(-5h), PM2.5(-5h), PM1(-5h), RH(-5h), T(-5h), PM10(-6h), PM2.5(-6h), PM1(-6h), T(-6h), PM10(-7h), PM1(-7h), T(-8h), RH(-11h), T(-11h), T(-12h)

It is evident that most of the chosen variables are PM concentrations and the only real time variable that was chosen for this system is Temperature. All of the other chosen variables are lagged measurements.

The performance differences of the models are presented in Table 1.

Algorithm/device	*MAE*	*Mean REU*	*R^2*	*% < 50% Uncertainty*
Feedforward NN	1,863	0,167	0,908	96,34
Random Forest	1,846	0,040	0,900	99,77
LSTM NN	1,831	0,042	0,955	100

Table 1: Table of the performance indicators (Mean absolute Error, Mean Relative Expanded Uncertainty, Coefficient of determination and Percentage of predictions under the 50% limit of Uncertainty) of the models for the Feedforward neural network, Random Forest algorithm and Long Short-term Memory network.

The MAE for the LSTM model is the lowest but only by a very small margin compared to the others. The FFNN model exhibits the highest mean REU, with the RF and the LSTM being the better performing algorithms in that respect. It is also remarkable that, according to the R^2 values, more than 90% of the targets' variance is captured by all the models. The best performing one is the LSTM NN, with a value of 0.955.

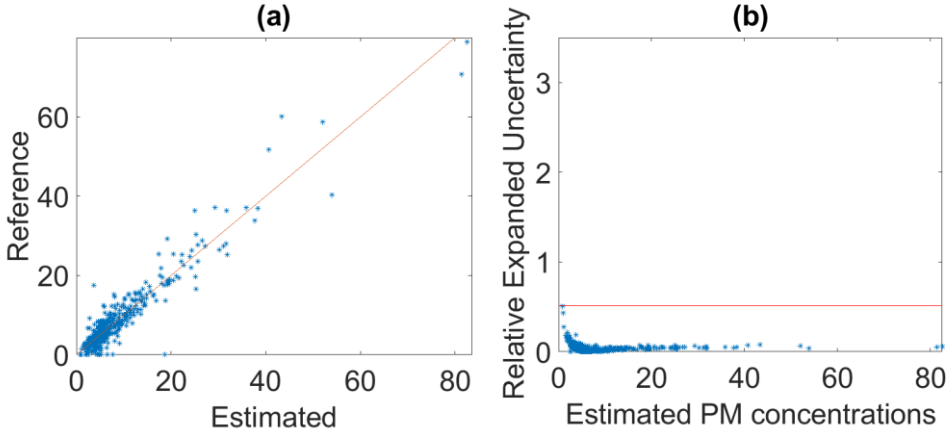


Fig. 2: (a) Regression plot of the estimated and reference PM10 concentrations for the LSTM NN model with respect to the y=x line (red line). (b) Graph of the Relative Expanded Uncertainty of every value that is estimated by the LSTM NN model. The red line represents the 50% uncertainty limit.

It should be underlined that the LSTM model presents a 100% of PM10 concentration predictions that fall under the 50% limit of uncertainty. This is important because it means that the model's performance indication is in accordance with the European Commission's directives.

4. Discussion

According to the results of this study, computationally improving LCAQS measurements is feasible and can be proven valuable for air quality monitoring.

If a final choice should be made, in order for one of those models to be deployed as part of an integrated solution for air quality monitoring, then it would be the LSTM NN algorithm. Its consistency and strong capability of accurately following the reference data with very low uncertainty is very difficult to ignore. Additionally, this machine learning algorithm's unique learning mechanism proves to be very effective in time series data because the model adapts on how to utilize past information for its predictions. However, it should be noted that there was a shortage of data for extensively testing the performance of the models and thus the results may be slightly different after multiple test sessions.

References

[1] Via: www.who.int

[2] Kumar P., Morawska L., Martani C., Biskos G., Neophytou M., Di Sabatino S., Bell M., Norford L., Britter R. (2015) : The rise of low-cost sensing for managing air pollution in cities. Published in Environment International 75, 199-205.

[3] Arlot S., Celisse A. (2010): A survey of cross-validation procedures for model selection. Published in Statistics Surveys 4, 40-79.

[4] Bergmeir C., Hyndman R. J., Koo B. (2018): A note on the validity of cross-validation for evaluating autoregressive time series prediction. Published in Computational Statistics & Data Analysis 120, 70-83.

[5] Hauke, J., Kossowski T. (2011): Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. Published in Quaestiones Geographicae 30(2), 87-93.

[6] Babatunde O. H., Armstrong L. , Leng J., Diepeveen D. (2014): A Genetic Algorithm-Based Feature Selection. Published in International Journal of Electronics Communication and Computer Engineering 5, 889-905.

[7] Breiman L. (2001): Random Forests. Published in Machine Learning 45, 5–32.

[8] Haykin S. S. (2009): Neural Networks and Learning Machines, Pearson Prentice Hall.

[9] Breiman L., Friedman J. H., Olshen R. A., Stone C. J. (1984): Classification and regression trees. CRC press.

[10] Hochreiter S., Schmidhuber J. (1997): Long short-term memory. Published in Neural computation 9(8), 1735-1780.

[11] Kvalseth T. O. (1985): Cautionary Note about R2. The American Statistician 39(4), 279-285.

[12] Borowiak A., Gerboles M. (2019): Air Quality Directive: Data quality requirements. Presentation at the Workshop Setting standards for low-cost Air Quality sensors, BAM, Berlin, http://netmon.eurice.eu/

[13] European Commission (2010): Guide to the Demonstration of equivalence of ambient air monitoring methods (2010). Via: https://ec.europa.eu/environment/air/quality/legislation/pdf/equivalence.pdf

[14] Via: https://www.mathworks.com/help/deeplearning/ug/long-short-term-memory-networks.html

PART IV
RECYCLING AND PLASTICS

Mechanical Recycling Considerations for Responsible Plastic Innovation

James Drayton[1], Justice Wright[2], Minh Nguyen Vo[3]

1. Introduction

In 1862, Alexander Parkes introduced the first successful cellulose plastic to the world, Parkesine [1]. Ever since, plastics have been remodeled and restructured to be stronger, more lightweight, and more flexible. However, the majority of plastics that contribute to today's widespread use are not composed of cellulose, but instead of synthetic polymers such as polyethylene terephthalate (PET); it has been estimated that bottles made from PET can take up to 450 years to decompose [2]. Beyond plastic bottles, plastics are used in our everyday life from to-go food packaging to the commercial waste bins used to dispose of our trash. Globally, we accumulate 275 million tonnes of plastic waste per year correlating to the 270 million tonnes of plastic produced per year [3]. Because not all plastic wastes are properly managed, the excess is carried into the following year's accumulation, which has resulted in global efforts to address the growing problem of plastics in the environment, especially single-use products.

There are several options for plastic disposal, and most are not considered environmentally friendly. Figure 1 illustrates some of the potential fates of consumer plastics. For mass disposal, most plastics are dumped into landfills where the plastic pile might be unable to decompose for up to 1,000 years according to some estimates [2]. Mass disposal also takes place through incineration which is expensive, pollutes the environment (notably through air emissions), and can harm the health of people if they are exposed to these emissions over time. On smaller scales, some plastics are subjected to mechanical recycling, in which they are melted down or ground up without chemically altering the polymer backbone to produce raw plastic that can be reused in new products. Plastics may also undergo chemical recycling, where they are broken back down into their component monomers and then re-synthesized into new products. Compared to disposal methods, both mechanical and chemical recycling save energy and conserve natural resources such as water and timber, though mechanical recycling is more efficient than chemical recycling [4].

[1] Argonne National Laboratory, Lemont, IL, United States, jdrayton@anl.gov
[2] Argonne National Laboratory, Lemont, IL, United States, justice.wright@anl.gov
[3] Argonne National Laboratory, Lemont, IL, United States, mvo@anl.gov

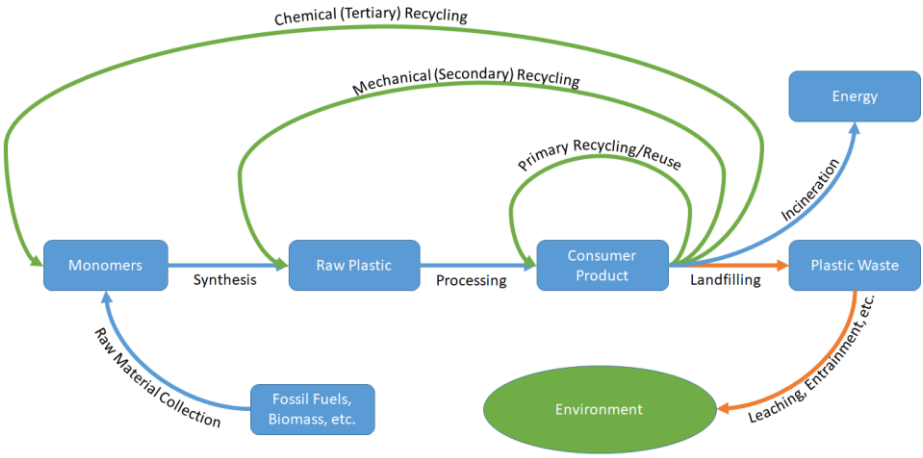


Fig. 1. Life cycle of consumer plastics.

Mechanical recycling begins when plastics are first collected from communities, either curbside or by delivery, and sent to a facility where they are separated by polymer category to be repurposed. While much of plastic sorting was previously done by hand, modern sorting facilities have added a variety of technological methods including Fourier-transform near-infrared spectroscopy to sort plastics by type, optical analysis to sort materials by color, and X-ray analysis for certain chemically-distinct plastics such as polyvinyl chloride (PVC). After sorting, plastics are typically ground to small flakes or pellets and cleaned to remove food waste and other contaminants before undergoing additional sorting, including sink/float separation based on density, froth flotation, additional spectroscopy techniques, and laser-sorting [5]. It has been concluded that the more procedures a facility has for sorting their plastics, the better the recovery of reusable plastic is and the greater the reduction in the use of recovery incineration.

Compared to feedstock recycling, engineered landfills, and energy recovery through incineration, mechanical recycling is preferred for waste management as it satisfies the three Rs of reduce, reuse, and recycle. Mechanical recycling reduces the amount of virgin feedstock required in plastic production, resulting in an overall reduction in energy requirements. However, use of recycled plastic in this manner is hindered by the degradation of the plastic during the recycling process. This degradation is often poorly understood by plastic product manufacturers, limiting adoption of recycled plastics in favor of virgin materials [6]. To combat this, it is important to understand the effects of recycling on commonly used plastics. Knowledge of these degradation effects and mechanisms can also inform the design of new materials with the objective of limiting degradation during mechanical recycling. In service of these goals, we aimed to examine current knowledge of recycling-induced degradation and to identify measures that would allow for reliable and useful characterization of current and future recycled materials.

2. Methods

Data on recycling-induced aging in a variety of polymers was collected between June and July 2020 from papers identified from searching five databases, including ScienceDirect, Wiley Online Library, and

Argonne's library system. Keywords used in the search included "mechanical recycling," "secondary recycling," "thermal aging," and "photo-oxidative aging." All articles examined were written in English.

3. Results

3.1. Data Organization

The literature search resulted in 35 articles published between 1989 and 2020 being selected and reviewed. Of these, 13 were examined in detail, and tabular data were extracted from 12. The articles examined contained 21 metrics to quantify polymer aging induced by plastic use and recycling. These metrics describe three categories of transformation: mechanical changes, chemical changes, and morphological changes. Figure 2 shows the metrics corresponding to each category.

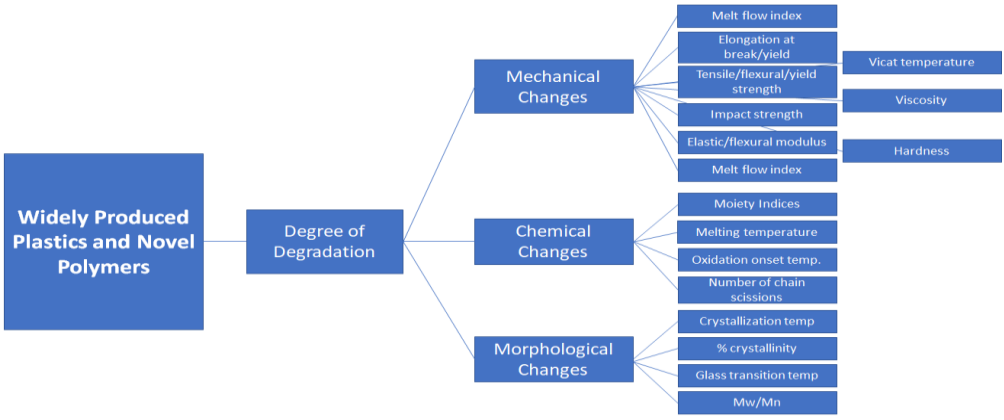


Fig. 2. Organization of polymer aging metric data.

Of the studies examined, 11 considered at least one metric of mechanical change, nine considered at least one metric of chemical change, and six considered at least one metric of morphological change. Within these categories, each study selected different metrics to investigate; the incidence rate of each metric in the examined studies is shown in Figure 3.

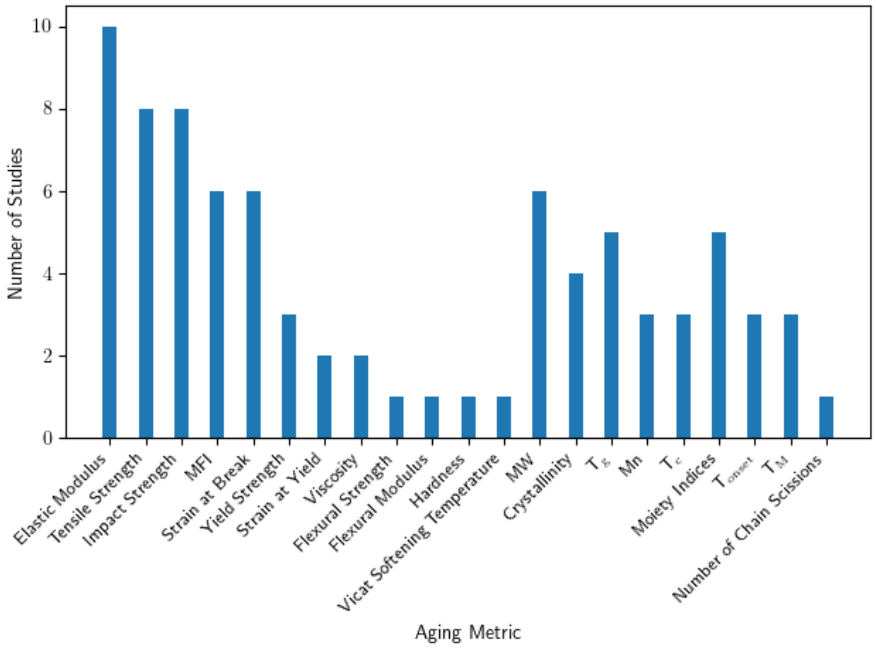


Fig. 3. Incidence of polymer aging metrics among studies examined.

Elastic modulus was the most commonly measured metric, discussed in 80% of studies examined. Tensile strength, impact strength, melt flow index (MFI), strain at break, and molecular weight were also common, with each represented in at least half the examined studies. Other metrics were much less common.

Many of these metrics can be measured in accordance with numerous standard methods. For example, MFI can be measured at either 190°C or 230°C, under various loads, including 2.16 kg or 5 kg (e.g., in accordance with the International Organization for Standardization [ISO] Standard 527); impact strength can be measured with or without a notch in the material and under a range of forces (e.g., in accordance with ISO 179). Further, virgin plastics encompass a wide variety of material characteristics, which can lead to a large numerical disparity in post-recycling metrics. As a result, measurements are often not directly comparable across studies or materials. To enable better comparison between the effects of recycling on different materials, metrics in this work are reported in terms of percent change from virgin material. Metrics for several polymers of interest are presented in the following sections.

3.2. PET

Polyethylene terephthalate (PET) is a widely used thermoplastic, often used in the beverage industry to create bottles. Because of its widespread use, it is an important target for recycling [7]. Measures of post-recycling aging in PET are summarized in Table 1.

The largest changes from recycling can be observed in the mechanical and morphological properties of PET. After one reprocessing cycle, the plastic already exhibits a large increase in degree of crystallinity, which continues to increase with more cycles. This increase in crystallinity is accompanied by sharp decreases in elongation at break, impact strength, and eventually tensile strength, as well as an increase in elastic modulus. These large property changes indicate that recycled PET is unsuitable for the same applications as virgin PET without extensive stabilization using additives, which may bring their own environmental risks, or complex reprocessing methods. Though applications for recycled PET do exist, it is attractive to develop alternatives to PET that exhibit lower degradation during recycling to enable their reuse in their original function.

Number of Reprocessing Cycles	Change from Virgin Material [%]				
	Tensile Strength	Elastic Modulus	Elongation at Break	Impact Strength	Degree of Crystallinity
1[7]	--	1.4	-98.0	-20.0	30.0
1[7]	--	-6.7	-98.9	-40.0	60.0
1[8]	-5.5	-21.5	-16.7	-47.4	96.7
2[8]	111.0	-11.2	-86.4	-81.5	107.4
3[8]	22.4	26.9	-93.8	-88.1	134.2
4[8]	-24.1	24.6	-96.2	-95.6	135.6
5[8]	-57.8	23.1	-98.3	-95.6	133.6

Table 1. Aging metrics in recycled PET.

These data also illustrate a major challenge to widespread adoption of recycled plastics in that the exact extent to which the aging metrics changed after one reprocessing cycle varied greatly between batches. This may be due to differences in the content of the pre-recycling plastics, deviations in the reprocessing temperature, differences in the equipment used, or a combination of these factors. While the general trends are similar enough across batches to be inferred from these samples, reliable prediction of the exact properties of recycled PET will require more extensive testing in a variety of conditions.

3.3. PLA

Polylactic acid (PLA) is a biopolymer derived from lactic acid. One of the most common biopolymers, it is often employed in 3D printing, among other applications [9]. Like PET, PLA is an attractive target for recycling due to its relatively widespread use. Measures of post-recycling aging in PLA are summarized in Table 2.

Number of Reprocessing Cycles	Change from Virgin Material [%]				
	Tensile Strength	Elastic Modulus	Elongation at Break	Melt Flow Index	Degree of Crystallinity
1[10]	-11.3	-5.3	--	--	--
1[9]	-1.2	-1.0	-50.0	--	--
2[9]	-8.8	-0.5	-41.7	--	--
3[9]	-14.3	-1.2	-46.7	--	--
4[9]	-32.8	0.5	-61.7	--	--
5[9]	-48.2	0.6	-70.0	--	--
5[11]	--	-3.8	--	71.7	-78.9

Table 2. Aging metrics in recycled PLA.

PLA exhibits relatively small changes in mechanical properties in the first three reprocessing cycles, with the exception of elongation at break, which decreases sharply after the first cycle but remains stable for several cycles thereafter. By the fifth reprocessing cycle, PLA undergoes more extensive mechanical degradation, indicated by a larger decrease in tensile strength, accompanied by severe morphological changes reflected in a sharp drop in degree of crystallinity. The late onset of this degradation indicates that PLA may be a reasonable choice for closed-loop recycling in service of a circular economy model, although it could not be used indefinitely in this manner and would eventually have to be redirected to a different application.

3.4. ABS

Acrylonitrile butadiene styrene (ABS) is a thermoplastic comprised of polybutadiene rubber mixed with styrene-acrylonitrile copolymer [12]. ABS is an interesting recycling target due to its widespread use across many industries, including applications in automobiles, communications, and electronic devices. Measures of post-recycling aging in ABS are summarized in Table 3.

Number of Reprocessing Cycles	Change from Virgin Material [%]				
	Tensile Strength	Elastic Modulus	Impact Strength	Degradation Onset Temperature	Glass Transition Temperature
1[13]	-7.9	18.8	-28.6	--	-5.9
1[12]	8.0	6.1	-40.6	1.8	-1.0
2[12]	7.1	-0.5	-34.4	2.3	1.0
3[12]	8.7	3.8	-44.4	1.5	-1.9
4[12]	6.4	-0.4	-40.0	2.3	-1.9
5[12]	8.9	2.5	-56.3	1.5	-1.9
10[12]	10.7	-0.3	-76.3	2.1	-2.9

Table 3. Aging metrics in recycled ABS.

ABS exhibits generally high resistance to recycling-induced aging; most properties remain stable even after 10 reprocessing cycles. The notable exception is impact strength, which decreases drastically after the first cycle and continues to decrease further after the fifth cycle. Nonetheless, ABS is a good candidate for closed-loop recycling in applications that do not require high impact strength, and could be reused extensively in these applications to minimize the amount of new ABS that must be synthesized.

4. Discussion

Previous work has identified three key factors contributing to the performance of polymers subjected to mechanical recycling: degree of degradation, presence of low molecular weight (LMW) compounds, and degree of mixing with other polymers. The data reviewed in this work are intended to quantify the degree of degradation. In conjunction with similar data quantifying the presence of LMW compounds and degree of mixing, these data can be used to identify polymers with poor performance in mechanical recycling and assess potential replacements. Aggregation of data describing the degree of degradation will be particularly useful in this effort because while the other broad factors depend mostly on the particulars of the recycling process, degree of degradation depends primarily on the material properties of the plastic being recycled. Thus, knowledge of the degradation performance of current widely used polymers can guide efforts to design new materials with similar material properties to those plastics that perform well in recycling. If degradation can be quantified such that the material properties of a recycled plastic can be reliably predicted, these data could also assist in identifying alternative uses for recycled plastics that are unsuitable for their original use. Future efforts should be directed towards expanding the number of materials for which degradation data are available, as well as towards understanding the underlying chemical factors in a polymer's recycling performance, to better enable this type of analysis.

Of the three subtypes of data examined, mechanical properties were the most commonly studied, and individual metrics of mechanical change such as elastic modulus and tensile strength were found amongst a majority of the studies. Knowledge of these properties is likely of particular interest to manufacturers, as it has a direct impact on the applications for which a recycled plastic is suited. However, coverage of chemical and especially structural changes was less widespread; further, the metrics used to quantify these changes were less standardized. These fields will be of equal or greater importance to chemists and engineers looking to design or select more recyclable materials. As such, it will be important for research to be directed towards understanding chemical and structural degradation during mechanical recycling.

The data reviewed in this work, and other similar data, will serve of use as part of a larger database to catalog the environmental implications of various plastics. With mechanical recycling as the most environmentally friendly method of returning the plastic to use, those materials with better performance can be considered to have comparatively lower environmental impact than those with poorer performance. Inclusion of these data in a database would also enable consideration of the environmental effects of additives used to improve post-recycling properties. Further, these data can be employed in conjunction with information on recycling stream compositions and quantities to suggest potential areas for improvement in the plastic recycling industry. If plastics with good recycling performance were found to be underrepresented in recycling streams, it would be attractive to take measures to increase the rate at

which those materials are recycled. The next phase of this work will consist of integrating the data reviewed in this study into an early concept of such a database.

In summary, this work identifies 21 metrics across three categories that can be used to quantify mechanical recycling-induced degradation in common and novel plastics. Currently, mechanical changes are better documented than chemical and structural changes. The authors recommend continued research into these areas. Eventually, the data reviewed here and similar data can be used to inform material selection and application in service of a circular economy.

Acknowledgements

The authors gratefully acknowledge Cristina Negri, Director of the Environmental Science Division at Argonne National Laboratory, for her vision in developing and implementing this research project. We also thank Young-Soo Chang for his technical review of this work. We further gratefully acknowledge Alicia Lindauer, Andrea Bailey, Nichole Fitzgerald, and additional colleagues in the U.S. Department of Energy, Office of Science, Bioenergy Technologies Office, for their support and funding for this research. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. http://energy.gov/downloads/doe-public-accessplan.

References

[1] Bellis, M. (2020): A Brief History of the Invention of Plastics. ThoughtCo (28 Jan.). www.thoughtco.com/history-of-plastics-1992322#:~:text=The%20first%20man%2Dmade%20plastic,retained%20its%20shape%20when%20cooled.

[2] LeBlanc, R. (2019): The Decomposition of Waste in Landfills. The Balance Small Business (22 Oct.). www.thebalancesmb.com/how-long-does-it-take-garbage-to-decompose-2878033.

[3] Ritchie, H. (2018): FAQs on Plastics. Our World in Data (2 Sept.). ourworldindata.org/faq-on-plastics.

[4] Cosate de Andrade, M.F.; Souza, P.M.S.; Cavalett, O.; Morales, A.R. (2016): Life cycle assessment of poly(lactic acid) (PLA): Comparison between chemical recycling, mechanical recycling and composting. Journal of Polymers and the Environment. 24, 372–384. doi:10.1007/s10924-016-0787-2.

[5] Hopewell, J.; Dvorak, R.; Kosior, E. (2009): Plastics recycling: challenges and opportunities. Philosophical Transactions of the Royal Society B: Biological Sciences. 364, 2115–2126. https://doi.org/10.1098/rstb.2008.0311.

[6] Vilaplana, F.; Karlsson, S. (2008): Quality Concepts for the Improved Use of Recycled Polymeric Materials: A Review. Macromolecular Materials and Engineering. 293, 274–297. doi:10.1002/mame.200700393.

[7] Torres, N.; Robin, J.J.; Boutevin, B. (2000): Study of thermal and mechanical properties of virgin and recycled poly(ethylene terephthalate) before and after injection molding. European Polymer Journal. 36, 2075–2080. doi:10.1016/S0014-3057(99)00301-8

[8] López, M. del M.C.; Ares Pernas, A.I.; Abad López, M.J.; Latorre, A.L.; López Vilariño, J.M.; González Rodríguez, M.V. (2014): Assessing changes on poly(ethylene terephthalate) properties after recycling: Mechanical recycling in laboratory versus postconsumer recycled material. Materials Chemistry and Physics. 147, 884–894. doi:10.1016/j.matchemphys.2014.06.034.

[9] Cruz Sanchez, F.A.; Boudaoud, H.; Hoppe, S.; Camargo, M. (2017): Polymer recycling in an open-source additive manufacturing context: Mechanical issues. Additive Manufacturing. 17, 87–105. doi:10.1016/j.addma.2017.05.013.

[10] Anderson, I. (2017): Mechanical properties of specimens 3D printed with virgin and recycled polylactic acid. 3D Printing and Additive Manufacturing. 4, 110–115. doi:10.1089/3dp.2016.0054.

[11] Shojaeiarani, J.; Bajwa, D.S.; Rehovsky, C.; Bajwa, S.G.; Vahidi, G. (2019): Deterioration in the physico-mechanical and thermal properties of biopolymers due to reprocessing. Polymers. 11, 58. doi:10.3390/polym11010058.

[12] Pérez, J.M.; Vilas, J.L.; Laza, J.M.; Arnáiz, S.; Mijangos, F.; Bilbao, E.; León, L.M. (2010): Effect of reprocessing and accelerated weathering on ABS properties. Journal of Polymers and the Environment. 18, 71–78 . doi:10.1007/s10924-009-0154-7.

[13] Brennan, L.B.; Isaac, D.H.; Arnold, J.C. (2002): Recycling of acrylonitrile–butadiene–styrene and high-impact polystyrene from waste computer equipment. Journal of Applied Polymer Science. 86, 572–578. doi:10.1002/app.10833.

Engineering for a Circular Economy: Key Factors for the Design of Biodegradable Plastics and Plastic-Degrading Enzymes

Mary Rommer[1], Margaret MacDonell[2]

1. Introduction

Since the 1950s, industrial societies have become increasingly dependent on synthetic polymers – specifically plastics – as their light weight, low cost, and durability are unmatched by other materials. By 2018, worldwide annual plastic production had reached 360 million metric tons [1], and annual production was projected to double by 2040 as population and demand increase [2]. Of the nearly 4 million metric tons of plastic produced in 2018, 63 million were improperly disposed of, thus serving as a potential source of accumulation in natural settings, including the world's oceans [1]. Plastic production and accumulation in the environment historically share a direct relationship. Extrapolating into the future, unless something changes to slow and ultimately reverse this trend, it is anticipated that used plastics will continue to accumulate in the environment as production continues to grow.

Plastic accumulation in the world's oceans has already become a deadly phenomenon for marine life. Waste plastics are estimated to kill more than 1 million marine animals a year through entanglement, ingestion with physical interference leading to starvation, or ingestion of toxic additives and adsorbed toxins [3]. Via trophic transfer, marine plastics have also been found to affect various trophic levels through the food chain, up to and including humans.

Although microorganisms that produce enzymes capable of bonding to and creating energy from synthetic plastic substrates might eventually evolve at a scale sufficient to make an impact, the very slow process to reach this natural evolutionary state extends beyond the time likely to matter for humankind. More than 92% of produced plastic is considered non-biodegradable and the accelerating production of plastic has already greatly outpaced rates of its use/reuse, recycling, and incineration [4]. The remaining plastics are either landfilled or lost to the environment.

Social movements have made some gains in public awareness resulting in increased efforts to reuse and recycle plastics; however, these efforts are not nearly sufficient to offset the continuing introduction of discarded plastics to the environment. A recent study outlined the fate of plastic produced annually: annually, 31% remains in use, 14% is recycled, 14% is incinerated (or used for energy recovery), and 20% is placed in managed landfills. The remaining 21%, over 77 million metric tons a year, is either leaked or improperly disposed of into the environment [1]. These post-use fates are illustrated in Figure 1.

[1] Argonne National Laboratory, Lemont, IL, USA, mrommer@anl.gov
[2] Argonne National Laboratory, Lemont, IL, USA, macdonell@anl.gov

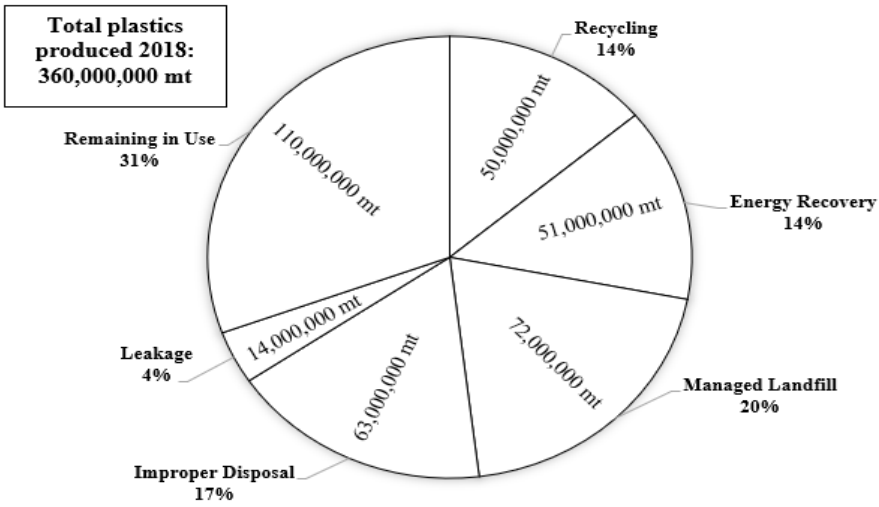


Fig. 1. The fate of plastics produced in 2018.

Many recycling and incineration facilities and managed landfills operate at relatively low environmental efficiencies and hinder efforts toward achieving a circular economy. For example, the quality of many plastics degrades during the recycling process, rendering them unsuitable for practical reuse, which contributes to high rates of single-use disposal or leakage. Incineration is not an environmentally preferred method of plastic disposal, with some estimating that this process releases 16 million metric tons of greenhouse gases each year, along with other potentially harmful chemicals [5]. Managed landfills continue to grow in volume, and associated leachate releases can adversely impact local environments.

To address the challenge of waste plastics, and to achieve environmental sustainability through a circular economy, industrial societies must either eliminate plastics – which play a significant role in daily life – or engineer materials, systems, plastic-degrading organisms, or all three such that the rate of plastic production is equivalent to the rates of use/reuse, recycling, energy recovery, and biodegradation. Scientists and engineers, seeking a solution to the 21% of current plastics being lost to the environment, are pursuing engineered bio-based materials that possess the performance attributes of synthetic plastics but are designed with an emphasis on key characteristics that enable increased reuse, recycling, energy recovery, and biodegradation. Such materials, in cooperation with genetically engineered enzymes capable of rapidly degrading polymers at an industrial scale, would have the potential to revolutionize industry. Engineering advances in areas such as renewable bio-based plastics and efficient plastic-degrading enzymes have the potential to propel the plastics industry into the realm of a circular economy.

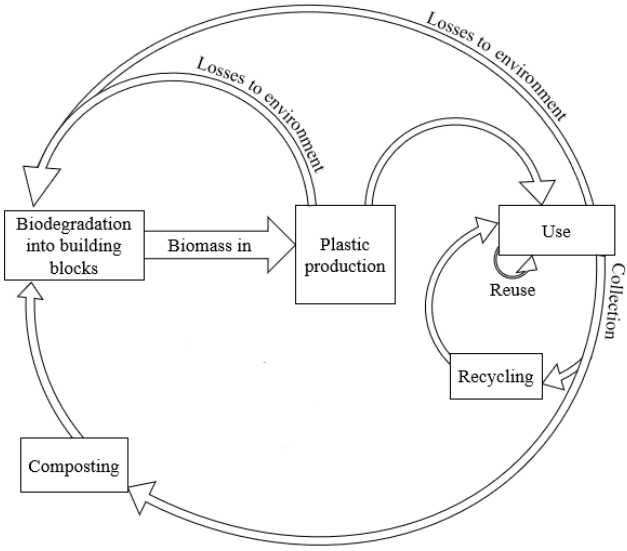


Fig. 2. Circular economy model.

Biodegradable plastics and enhanced enzymes for plastic degradation are not novel concepts. Plastic biodegradation has been a topic of interest in the scientific world since the 1970s, yet limited progress has been made in the five decades since, due to multiple barriers. Not only must bio-based plastics maintain flexibility, light weight, and low cost, they must also remain durable enough for their intended purpose. Designing for durability throughout the useful life of a plastic product and for degradation after that use ends seems contradictory, yet scientists must overcome this barrier to create a circular plastic. Furthermore, it was not until 2016 that scientists discovered an enzyme that exemplifies enzymatic evolution for the degradation of plastic. The discovery of a naturally evolved plastic-degrading enzyme has enabled substantial progress in engineering enzymes tailored for this purpose, but much more work is needed before waste plastics can be efficiently degraded at an industrial scale.

In order to develop circular plastics and efficient plastic-degrading enzymes, it is necessary to: (1) outline the properties of plastics that affect both the plastic's durability and degradability; and (2) outline the properties of plastic-degrading enzymes that make them effective at this task. Once these properties are understood, engineering solutions can be developed and optimized to inform the design of biodegradable plastics and plastic-degrading enzymes.

Over the last few decades, many scientists have conducted studies to better understand the durability, degradability, and enzymatic degradation of plastics, and these studies have advanced the field in small steps toward a circular economy. What is needed is for key data relevant to the environmental fate of used plastics to be integrated and organized in a way that enables developers of a new polymer to consider its post-use fate during the design phase The purpose of the current study is to better understand the state of knowledge relevant to plastic biodegradation, and to lay the groundwork for an integrated database that combines polymer, disposition, and environmental fate data, also including a compilation of enzymes that can degrade plastics. A specific aim is to pinpoint key characteristics of plastics that affect degradation as

well as characteristics of key enzymes that promote degradation. These key characteristics may then be used to guide the engineering of circular economy materials and plastic-degrading enzymes. An early step toward this process is illustrated by considering a largely produced non-biodegradable plastic: polyethylene terephthalate (PET).

2. Methods

A scientific literature review was conducted, from May through July 2020 to identify online studies and peer-reviewed publications that relate to the biodegradability of existing plastics in the environment, as well as enzymes that can degrade plastics. This search was conducted through the Argonne National Laboratory research library and considers articles published in the English language from the year 2005 to 2020. Key search terms for the plastic focus included: bio-based, biodegradable, circular economy, biodegradation rate, production (tons), common/widely used, and accumulation. Key terms for the enzyme focus included: plastic degrading, enzymatic degradation, abundance, distribution, and growth conditions.

Data extracted from the literature are organized using spreadsheets, with one for the characteristics of plastics that affect their biodegradability and another – which can be cross-referenced with the first – for data on microorganisms effective for plastics degradation, the associated enzymes and effective mutants, and the specific plastics they can degrade. The intent is to expand the compilation as further studies are conducted and reviewed.

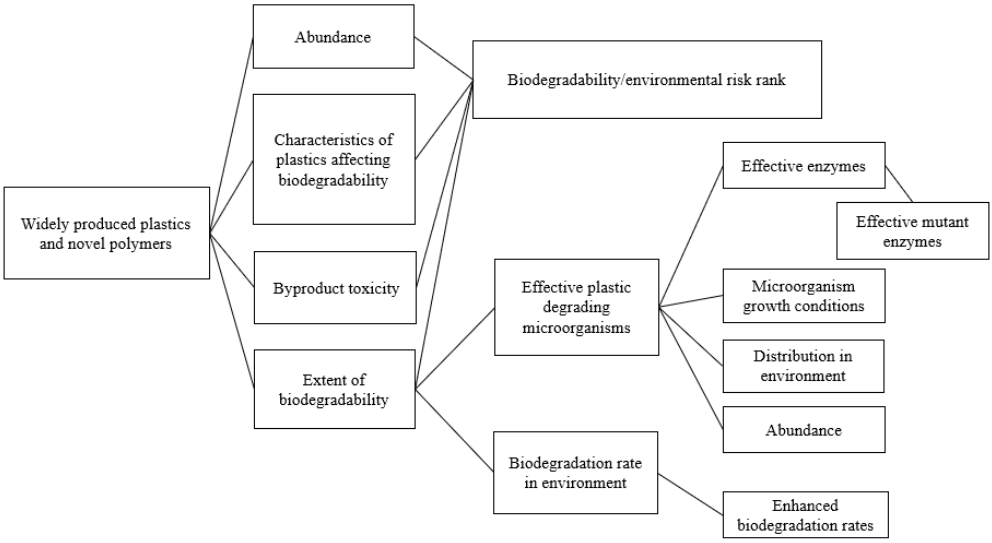


Fig. 3. Data schema for biodegradable plastics and plastic-degrading enzymes.

A further spreadsheet is being developed to enable ranking of the biodegradability and related environmental risk of various plastics, using a decision matrix. The ranking will depend on the characteristics affecting plastic biodegradation, abundance of each plastic type, toxicity of byproducts of biodegradation, and the extent of biodegradability of each plastic as determined from the biodegradation

rate, and the number, abundance, and distribution of effective degrading enzymes in the environment. As studies on enzyme mutations gain more traction, those data would further inform the ranking.

3. Results

3.1. Definitions

The literature search uncovered a variety of definitions for terms such as bioplastic, bio-based, degradable, biodegradable, non-biodegradable, and compostable. Selected definitions considered for the ongoing project are illustrated below.

- Bio-based: Constructed from renewable resources, such as plant matter; note that a bio-based origin does not necessarily indicate the plastic is biodegradable or naturally compostable.
- Biodegradable: Can be broken down by enzymes from naturally occurring microorganisms during a specified time frame and in a specified environment; the products or biodegradation include water (H_2O), carbon dioxide (CO_2) and/or methane (CH_4), biomass and inorganic compounds.
- Bioplastic: Either a bio-based or biodegradable plastic, or both.
- Compostable: Can be biodegraded in an industrial composting facility (at elevated temperatures) within a specified time frame.
- Degradable: Can be degraded over time. Note that all plastics can eventually degrade to some extent, including through physicochemical processes such as absorption of ultraviolet (UV) radiation (photodegradation), which breaks polymer bonds.
- Non-biodegradable: Cannot be biodegraded by enzymes from naturally occurring microorganism into natural organic and inorganic compounds at a substantial rate relative to a human time scale.

Widely used plastics and novel polymers identified from the online information resources reviewed include: polyethylene (PE, high density PE [HDPE]), low density PE (LDPE), linear low density PE (LLDPE), polyethylene terephthalate (PET), polypropylene (PP), polyvinyl chloride (PVC), polyurethane (PUR), polystyrene (PS), expanded PS (EPS), polyethylene furanoate (PEF), polylactic acid (PLA), thermoplastic starch (TPS), polyhydroxyalkanoate (PHA), polycaprolactone (PCL), and polybutylene adipate terephthalate (PBAT). Information compiled to date regarding the nature (fossil- or bio-based), biodegradability, and compostability of these plastics is summarized in Table 1.

	Non-biodegradable	Compostable	Biodegradable
Bio-based	PEF	PLA	TPS, PHA
Fossil-based	PP, PS (including EPS), PVC		PCL, PBAT
Fossil-based or bio-based	PE (including HDPE, LDPE, LLDPE), PET, PUR		

Table 1. Feedstock type and degradability of widely produced and novel plastics[1].

It is important to note that, while plastics are degradable, not all plastics are biodegradable. When plastics are released to the natural environment, e.g., as litter, they are commonly exposed to ultraviolet radiation. Impurities in the plastics absorb this radiation, which then excites photons in the plastics, creating free radicals. When oxygen is present, the free radicals – along with the oxygen – form compounds that break the polymer bonds holding plastic molecules to one another. As this process continues over a relatively long time, it can eventually break a discarded plastic product into microplastics and nanoplastics. While these plastic particles are smaller than the original plastic they made up, their chemical properties are unchanged. Thus, the plastic molecules still possess various properties of plastics, which can potentially impact environmental health; for example, hydrophobicity can enhance the adsorption of chemical pollutants from the local environment. Degradation does not reduce the amount of plastic in the environment; however, it can help to accelerate biodegradation by increasing the surface area of the plastic substrate, allowing increased biofilm growth and activity [6].

Biodegradation of plastics is a process by which enzymes adhere to the plastic substrate, forming an enzyme-substrate complex. This binding and hydrolysis subsequently stress and break the glucose-fructose bonds within the plastics, thus breaking down the plastic molecules. Once significantly broken down, mineralization into organic constituents including water, carbon dioxide, methane, and biomass and inorganic compounds could occur. While additives might still pose an environmental hazard at certain levels, the plastic itself is commonly broken down into natural fate products. Biodegradation is thereby a desirable process for the natural mitigation of discarded plastics.

Compostable plastics are a step towards being more environmentally friendly, but these materials are only biodegradable under specifically controlled environments. Unfortunately, the literature indicates that most widely produced plastics (>92%) are non-biodegradable including HDPE, LDPE, and LLDPE (36% of plastics produced), PP (21% of plastics produced), PS and EPS (<10% of plastics produced), PVC (12% of plastics produced), PUR (<10% of plastics produced), and PET (<10% of plastics produced, not including fibers) [4]. It is necessary to address the large proportion of non-biodegradable plastics in order to achieve circularity of materials, which is the ultimate goal.

[1] Shading indicates information to be pursued during the next phase of the project.

3.2. Characteristics of Polyethylene Terephthalate (PET) Affecting Biodegradation

To illustrate the types of data available related to polymer biodegradation, PET was selected as an example plastic. The characteristics of PET that affect its biodegradability, and those of PET-degrading enzymes, are used to outline a preliminary structure and organization for this portion of the planned database. PET belongs to the polyester class of polymers and is a non-biodegradable, non-compostable plastic that can be either bio-based (bio-PET) or derived from fossil resources. PET and bio-PET consist of alternating subunits of terephthalic acid (TPA) and ethylene glycol (EG), linked by ester bonds [6]. Traditional PET precursors, purified terephthalic acid (PTA) and EG, are produced from fossil resources. Bio-PET, however, is produced from bio-based PTA and fossil-based EG, fossil-based PTA and bio-based EG, or both bio-based PTA and EG [19]. The ratios of bio-based to fossil-based material vary along with the percent of biomass as seen in Table 2 (adapted from [19]).

PTA feedstock	EG feedstock	% Biomass
Fossil	Fossil	0
Fossil	Corn, switchgrass, or wheat straw	30
Wood	Fossil	70
Wood	Corn, switchgrass, or wheat straw	100
Corn stover	Fossil	70
Corn stover	Corn, switchgrass, or wheat straw	100

Table 2. Fossil- and bio-based PET feedstock combinations and ratios.

Although bio-PET is gaining interest in industry, the majority of TPA and EG monomers used to construct PET are still fossil-based [7]. While bio-PET does not surpass PET in production rates, environmental factors should be considered when examining the environmental friendliness of bio-PET upon natural degradation. Multiple environmental considerations are involved across the life cycle of a plastic, from creation to degradation. These include carbon dioxide contributions, fossil resource depletion, acidification, terrestrial eutrophication, human health effects, ecotoxicology, smog, and ozone depletion. As seen in Figure 4 (adapted from [19]), partial and full bio-based PET "have worse performance than their 100% fossil-based counterparts" [19].

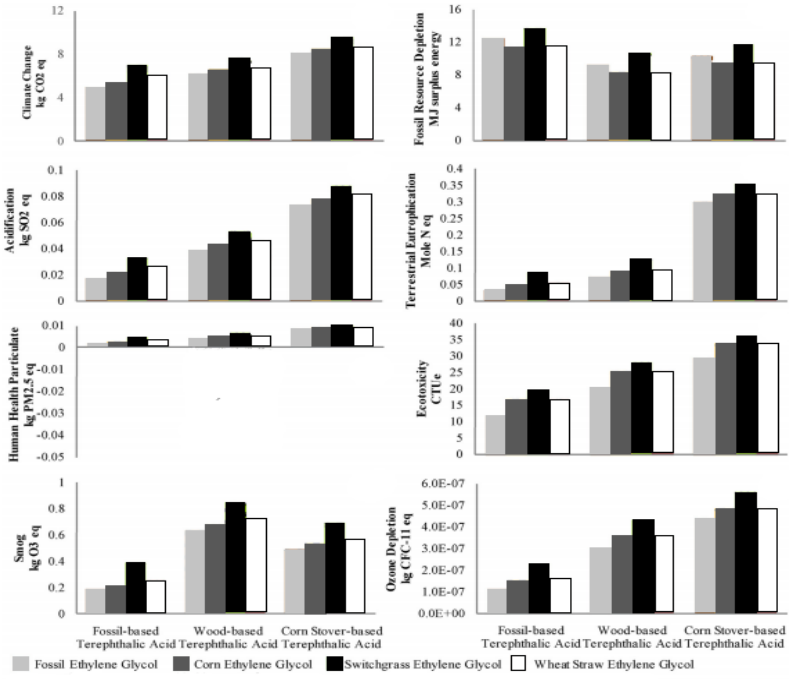


Fig. 4. Enzymatic degradation of PET under various reaction conditions.

Despite a push towards bio-friendly and bio-based plastics, the terms are not synonymous. The data in Figure 4 indicate that bio-based plastic is not necessarily bio-friendly. Furthermore, bottle-grade (high-crystallinity) PET and bio-PET exhibit a similar recalcitrance to biodegradation and recycling due to their similar composition of high ratios of aromatic compounds, despite the source of said compounds, which make them chemically inert [12].

PET is a high-molecular-weight thermoplastic that can be amorphous and semi-crystalline. More than 49 million tons of PET fibers were produced worldwide in 2014, and globally, PET resin production reached nearly 28 million tons in 2015 [7]. Because PET is widely used in single-use packaging for food and beverages and is thus prone to improper disposal (e.g., littering), understanding its potential for biodegradation is an active research topic. Although some efficient plastic-degrading enzymes have been identified for low-crystallinity PET, in general, PET is not classified as a biodegradable plastic because the rate of biodegradation is not sufficient for significant high-crystallinity PET degradation within a reasonable time frame. Nevertheless, slow biodegradation has been observed in numerous studies.

A number of factors have been identified as affecting the biodegradability of plastics, including the hydrophobicity, surface topography, crystallinity, and molecular size of the polymer [8]. For PET biodegradation, fluctuation of the polymer chain is also important, specifically at temperatures higher than the glass transition temperature (Tg), which is about 75°C to 80°C for PET [9,10]. Polymer chain fluctuation exposes the chain to enzymatic attack and is also affected by the water absorbency of PET,

which ranges from 0.1 to 1.0% [9]. The greater the water absorbency, the more flexible the polymer chain, and water absorbency increases with temperature and exposure to water. When water molecules enter the polymer chain, they increase chain flexibility by weakening hydrogen bonds and randomizing the chain. Increased PET degradation rates are thus associated with diffusion of water molecules into the polymer chain and reaction temperatures higher than Tg which allow even the brittle portions of PET to exhibit increased susceptibility to enzymatic attack [11].

The crystallinity of PET varies with the product. Most PET bottles are of high crystallinity, ranging from 30-40% [9]. The PET used in packaging is typically more amorphous, with only approximately 8% crystallinity [9]. Increased crystallinity limits the fluctuation of polymer chains, thus decreasing their exposure to enzymatic degradation. The amorphous region of a plastic is likely the first region attacked due to its enhanced flexibility. The degradation of this amorphous region increases the susceptibility of the crystalline region to degradation. Therefore, to increase plastic strength properties, it is common for PET bottles to be stretched such that the polymer chains are oriented to strengthen the bonds between molecules by increasing the crystallinity and reducing the percentage of amorphous material. This in turn greatly limits the susceptibility of PET to degradation such that no efficient degrading enzymes have yet been identified. Factors that affect the flexibility of the PET polymer chain thus affect the biodegradability of the inner PET block. This biodegradability also depends on the availability of enzymes capable of binding to the PET substrate and their ability to metabolize PET. Characteristics that affect the biodegradability of PET are highlighted in Table 2.

Plastic	**Polyethylene Terephthalate (PET)**
Chemical formula	$(C_{10}H_8O_4)_n$
Aromatic compound content	High, chemically inert
Biodegradability	Non-biodegradable
Feedstock	Fossil resources (highest production), bio-based
Annual production	49.2 million tons PET fibers in 2014, 27.8 million tons PET resins in 2015
Use	Single use packaging, bottles
Average crystallinity	Bottle grade: 30-40%, packaging: ~8%
Molecular weight (g/mol)	8,000-31,000
Molecular weight of repeat unit (g/mol)	192.2
Crystalline density at room temperature (g/cm^3)	1.50
Amorphous density at room temperature (g/cm^3)	1.33
Glass transition temperature (°C)	~75-80
Melting temperature (°C)	250
Thermoplasticity	Thermoplastic
Hydrophobicity	Hydrophobic
Water absorbency	0.1-1.0%

Table 3. Characteristics of PET that affect its biodegradability.

3.3. Enzymatic Degradation

As plastic becomes an increasingly abundant substrate in the environment, it is likely that microorganisms will eventually evolve that can degrade this material and use the byproducts for energy. In fact, we can already observe this slow process in action. Microorganisms – most commonly including the bacteria and fungi taxonomies [18] – have evolved over billions of years to be capable of degrading and metabolizing cutin, the thin waxy layer found on plant leaves. These enzymes are known as cutinases. Because cutin is an aliphatic polyester, it is structurally similar to plastic in its bonding; for this reason, some cutinases are able to degrade certain plastics, albeit with limited efficiency. According to Kawai et al. [9], all identified PET hydrolases are cutinases or are expected to be capable of degrading cutin.

A PET-degrading microorganism was discovered in 2016 at a plastic bottle recycling factory in Sakai, Japan [12]. Named for this city, *Ideonella sakaiensis* illustrates the promise of evolution, and efforts are under way to accelerate this natural process. The enzymes from this organism that had adapted to a plastic waste environment were studied to assess their effectiveness as PET degraders, beginning with low-crystallinity, 1.9%, PET film as the main carbon source [12]. *I. sakaiensis* was found to form a biofilm over the PET substrate and extensively degrade the plastic. The microorganism consortium from which this species was isolated nearly degraded the entire PET film after 6 weeks at a controlled temperature of 30°C; surface pitting of the film was observed after incubation with *I. sakaiensis* proteins for 18 hours at this temperature. The rate of degradation was estimated to be 0.13 mg/cm^2/day. Furthermore, 75% of the PET film's carbon was converted to CO_2 at 28°C.

Microbial consortia that lacked *I. sakaiensis* were unable to degrade the PET. The enzyme produced by this microorganism – mono(2-hydroxyethyl) 201-F6, termed PETase – was found to degrade PET as an energy and carbon source, and PETase is capable of enzymatically converting PET to mono-2-hydroxyethyl terephthalate (MHET), bis(2-hydroxyethyl) terephthalate (BHET), and trace amounts of TPA. Another enzyme secreted by *I. Sakaiensis*, MHETase, can further break down the plastic into monomers TPA and EG. The effectiveness of PETase on high-crystallinity PET is not yet known.

The genome sequence of *I. sakaiensis* was examined to explore the genes that enable PET hydrolysis. It was reported that the "one identified open reading frame (ORF), ISF6_4831, encodes a putative lipase that shares 51% amino acid sequence identity and catalytic residues with a hydrolase from *Thermobifida fusca* (TfH) … that exhibits PET-hydrolytic activity" [12]. Other studies have also identified *T. fusca* as a capable PET degrader, beginning in 2005 [13]. This study found that the inner block of PET can be "effectively depolymerized by a hydrolase from the actinomycete *T. fusca"* and reported degradation at a rate of 8 to17 µm per week when incubated at 55°C. Many studies since have applied mutagenesis to cutinases such as that secreted by *T. fusca* to improve the PET-degrading efficiency under mild reaction conditions as well as other methods to increase the enzyme degradation capabilities.

Through a study of the crystal structure of the cutinase and substrate, informed mutations were introduced to the *T. fusca* cutinase TfCut2 [14]. Building on an earlier approach that attempted to increase PETase activity by adding surfactant [15], it was hypothesized that applying this approach to a stable cutinase such as TfCut2 would increase catalytic activity. This TfCut2 enzyme possesses a negatively charged surface, thus, a cationic surfactant was applied to create electrostatic attraction. This electrostatic

interaction between the surfactant and enzyme is thought to increase enzyme concentrations near the surface of the plastic, allowing increased hydrophobic interactions to accelerate biodegradation.

Alkyl trimethyl ammonium chloride was tested as a surfactant with varying chain lengths and under various reaction temperatures. The cationic surfactant was found to accelerate biodegradation of low crystallinity PET to a rate 13-fold faster than biodegradation in the absence of a surfactant. At all temperatures tested (40°C to 70°C), the addition of a surfactant greatly enhanced hydrolysis capabilities. The greater the reaction temperature, the lower the surfactant concentration needed – and the shorter the alkyl chains – to produce the same accelerated activity as that of higher concentrations of surfactant at lower temperatures [14]. Furthermore, it was concluded that cationic surfactant did not affect the long-term enzymatic hydrolysis process of TfCut2. Over time the amount of PET biodegradation continued to increase linearly [14].

Improved activity of the TfCut2 enzyme was also reported following targeted mutagenesis based on comparisons to PETase sequences. The mutant enzyme TfCut2 G62A was found to successfully exhibit enhanced degradation about twice that of the wild-type enzyme. Additionally, replacing F209 increased the activity by as much as 7 times. These mutagenic breakthroughs resulted in the double mutant TfCut2 G62A/F209A, which exhibited much higher hydrolytic activity than the normal enzyme. The mutated cutinase combined with surfactant degraded PET at a higher rate (31 ± 0.1 nmol min^{-1} cm^{-2}), nearly 13 times that of the wild type.

While this progress seems substantial, it is important to recognize that low-crystallinity PET can biodegrade relatively easily compared with bottle-grade PET, which may be as much as 40% crystalline. Low-crystallinity PET is widely implemented for single-use packaging (including food wrappers), and although the ultimate aim of this enzyme research is to achieve industrial-scale biodegradation, enhanced by surfactant addition, the biodegradation of highly crystalline PET by TfCut2 and other enzymes is still highly inefficient.

Others have sought to address the issue of high-crystallinity PET biodegradation through protein engineering, focusing on three proteins that affect the ability of the PETase enzyme to bind to and release the substrate: R61, L88, and I179 [16]. The amino acid R61 is involved in the binding of PETase to the substrate. Its modified form R61A, which is more hydrophobic and has a lower molecular weight, altered the charge surrounding the substrate binding groove, thus increasing contact between the PETase and PET. This mutation offered a modest 1.4-fold increase in PET degradation compared to the wild-type PETase.

A mutant form of the L88 amino acid also involved in binding PETase to PET, L88F, altered the hydrophobicity and enhanced the interaction of PETase and PET, more than doubling the degradation rate compared to the wild type. Another amino acid, I179, also affects the substrate-binding groove, and its mutated form also more than doubled the degradation activity, This study helped identify key amino acids affecting the PETase binding groove to inform the development of future PETase mutants with higher degradation capabilities [16].

Many other studies have also attempted to mutate PETase for enhanced degradation, including another that focused on the binding site of the PETase enzyme [17]. That analysis revealed that PETase "retains ancestral α/β-hydrolase fold but exhibits a more open active-site cleft than homologous cutinases." By a mutation of active-site residues to conserved amino acids, improved PET degradation was observed [17].

Still the degradability of PET is limited, despite small advancements that illustrate the potential for a slow natural evolution process toward enzymes that can degrade plastics at the end of their product life.

The properties of effective PET hydrolases have been summarized as follows [9]: (1) a shallow, open active site accessible to the solvent (open lid structure); (2) adequate space for aromatic compounds; (3) an active cleft affinity with hydrophobic substrates; and (4) thermostability at temperatures higher than the glass transition temperature. These factors are among the key data elements being compiled to inform the innovation of novel polymers that can be more readily biodegraded after their use life ends plastics and enhanced enzymatic degradation.

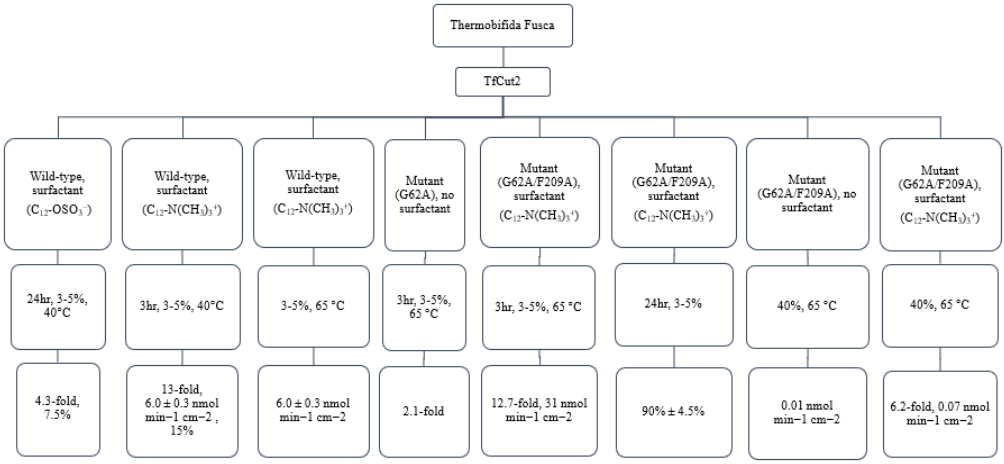


Fig. **5.** Enzymatic degradation of PET under various reaction conditions.

4. Discussion

The current project aims to develop a practical database that can help inform the design of more bio-friendly plastics. This paper focuses on one aspect of this database concept – biodegradation – and the literature evaluated has uncovered key data elements for incorporation into the planned database. Data relevant to this theme are limited and scattered, largely reflecting laboratory studies conducted under controlled conditions not found in the natural environment.

As the problem of environmental plastics continues to grow, it will be increasingly important to develop and integrate key data toward providing those who seek to develop more bio-friendly polymers a with practical information resource to inform that design. Within the biodegradation theme, this literature analysis has highlighted several opportunities for future research. One relates to the measurement method, with weight loss being a common metric for assessing plastic biodegradation. Unfortunately, this measure provides no information about whether or how much of the amorphous versus crystalline regions have changed. For this reason, simply measuring weight loss does indicate what part of the polymer has been degraded, which in turn limits practical interpretation for extrapolation to other plastics.

Another opportunity relates to enhanced enzymatic degradation. Scientists and engineers have successfully boosted the biodegradation rate for low-crystallinity plastics through mutant enzymes and surfactant additions, albeit at a very modest scale. The knowledge gained from these studies is anchoring follow-on explorations, including for more highly crystalline plastics. Better understanding of the biodegradability of these recalcitrant polymers could help inform the design of replacement products toward that could more readily biodegrade after their use life ends. No single approach can resolve the challenge posed by waste plastics in the environment. Rather, a multi-pronged effort is needed, to tap the knowledge being developed across all opportunity fronts, including regarding how to enhance polymer chain flexibility and better understand the relationship between this flexibility and enzymatic degradation following use while maintaining product durability throughout its design life.

The biodegradability information identified and synthesized as part of this project is being aligned with data that characterize other aspects of the overall problem. The data extracted for each theme are being integrated to outline the database concept. The plan is to provide a modular link to support lifecycle analysis that include the back end, after use and disposition, toward improving overall sustainability. The aim of this data outlining and integration is to promote the design of novel polymers that enable recyclability, reuse, energy recovery, bio-based plastics, and reduced use of additives that could pose a hazard upon release to the environment.

Biodegradability will remain a key theme, as discard and leakage of plastics are expected to continue for years until cost-effective replacements are widely available. That is, used plastics will continue to be released to the environment despite increased efforts toward recyclability, energy recovery, and improved efforts towards proper disposal in the industrialized world. PET was selected for illustration in this assessment because it is so widely used. Other types of plastics that will be similarly assessed include PE, HDPE, LDPE, LLDPE, PET, PP, PVC, PUR, PS, and EPS as well as novel polymers as they continue to evolve, such as PEF, PLA, TPS, PHA, PCL, and PBAT. The database being developed intends to ultimately serve a broad community of stakeholders interested in designing novel bio-based polymers that are more compatible with the environment throughout and beyond the product life.

Acknowledgements

The authors gratefully acknowledge Cristina Negri, Director of the Environmental Science Division at Argonne National Laboratory, for her vision in developing and implementing this research project. We also thank Young-Soo Chang for his technical review of this work. We further gratefully acknowledge Alicia Lindauer, Andrea Bailey, Nichole Fitzgerald, and additional colleagues in the U.S. Department of Energy, Office of Science, Bioenergy Technologies Office, for their support and funding for this research. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Minority Serving Institutions Partnership Program (MSIPP). The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display

publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. http://energy.gov/downloads/doe-public-accessplan.

References

[1] Conversion Market & Strategy. (2019): Global Plastics Flow 2018. https://www.euromap.org/files/Global_Plastics_Flow_Summary_Oct_2019.pdf?download=1.

[2] Lebreton, L.; Andrady, A. (2019): Future scenarios of global plastic waste generation and disposal. Palgrave Communications. 5, 6. https://doi.org/10.1057/s41599-018-0212-7.

[3] Sea Turtle Conservancy. (2020 accessed): Information About Sea Turtles: Threats from Marine Debris. https://conserveturtles.org/information-sea-turtles-threats-marine-debris/#:~:text=of%20Tracked%20Turtles-,Information%20About%20Sea%20Turtles%3A%20Threats%20from%20Marine%20Debris,in%20oceans%20around%20the%20world.

[4] Geyer, R.; Jambeck, J.; Law, K.L. (2017): Production, use, and fate of all plastics ever made. Science Advances. 3, e1700782. doi: 10.1126/sciadv.1700782.

[5] Global Alliance for Incinerator Alternatives (GAIA) (2020 accessed): The Hidden Climate Polluter: Plastic Incineration. https://www.no-burn.org/hiddenclimatepolluter/.

[6] Chamas, A.; Moon, H.; Zheng, J.; Qiu, Y.; Tabassum, T.; Jang, J.H.; Abu-Omar, M.; Scott, S.L. Suh, S. (2020): Degradation rates of plastics in the environment. ACS Sustainable Chemistry & Engineering. 8, 3494-3511. https://doi.org/10.1021/acssuschemeng.9b06635.

[7] Wei, R.; Zimmermann, W. (2017): Biocatalysis as a green route for recycling the recalcitrant plastic polyethylene terephthalate. Microbial Biotechnology. 10, 1302-1307. https://doi.org/10.1111/1751-7915.12714.

[8] Tokiwa, Y.; Calabia, B.P.; Ugwu, C.U.; Aiba, S. (2009): Biodegradability of plastics. International Journal of Molecular Sciences. 10, 3722-3742. doi: 10.3390/ijms10093722.

[9] Kawai, F.; Kawabata, T.; Oda, M. (2019): Current knowledge on enzymatic PET degradation and its possible application to waste stream management and other fields. Applied Microbiology and Biotechnology. 103, 4253-4268. https://doi.org/10.1007/s00253-019-09717-y.

[10] Scott, C. (2020): poly(ethylene terephthalate) information and properties. http://www.polymerprocessing.com/polymers/PET.html.

[11] Oda, M.; Yamagami, Y.; Inaba, S.; Oida, T.; Yamamoto, M.; Kitajima, S.; Kawai, F. (2018): Enzymatic hydrolysis of PET: functional roles of three Ca2+ ions bound to a cutinase-like enzyme, Cut190*, and its engineering for improved activity. Applied Microbiology and Biotechnology. 102, 10067-10077. doi: 10.1007/s00253-018-9374-x.

[12] Yoshida, S.; Hiraga, K.; Takehana, T.; Taniguchi, I.; Yamaji, H.; Maeda, Y.;, Toyohara, K.; Miyamoto, K.; Kimura, Y.; Oda, K. (2016): A bacterium that degrades and assimilates poly(ethylene terephthalate). Science. 351, 1196-1199. doi: 10.1126/science.aad6359.

[13] Müller, R.-J.; Schrader, H.; Profe, J.; Dresler, K.; Deckwer, W.-D. (2005): Enzymatic degradation of oly(ethylene terephthalate): Rapid hydrolyse using a hydrolase from T. fusca. Macromolecular Rapid Communications. 26, 1400-1405. https://doi.org/10.1002/marc.200500410.

[14] Furukawa, M.; Kawakami, N.; Tomizawa, A.; Miyamoto, K. (2019): Efficient degradation of poly(ethylene terephthalate) with *Thermobifida fusca* cutinase exhibiting improved catalytic activity generated using mutagenesis and additive-based approaches. Scientific Reports. 9, 16038. https://doi.org/10.1038/s41598-019-52379-z.

[15] Furukawa, M.; Kawakami, N.; Oda, K.; Miyamoto, K. (2018): Acceleration of enzymatic degradation of poly(ethylene terephthalate) by surface coating with anionic surfactants. ChemSusChem. 11, 4018-4025. https://doi.org/10.1002/cssc.201802096.

[16] Ma, Y.; Yao, M.; Li, B.; Ding, M.; He, B.; Chen, S.; Zhou, X.; Yuan, Y (2018).: Enhanced poly(ethylene terephthalate) hydrolase activity by protein engineering. *Engineering*, 4(6), 888-893. https://doi.org/10.1016/j.eng.2018.09.007.

[17] Austin, H.P.; Allen, M.D.; Donohoe, B.S.; Rorrer, N.A.; Kearns, F.L.; Silveira, R.L.; Pollard, B.C.; Dominick, G.; Duman, R.; El Omari, K.; Mykhaylyk, V.; Wagner, A.; Michener, W.E.; Amore, A.; Skaf, M.S.; Crowley, M.F.; Thorne, A.W.; Johnson, C.W.; Woodcock, H.L.; McGeehan, J.E.; Beckham, G.T. (2018): Characterization and engineering of a plastic-degrading aromatic polyesterase. Proceedings of the National Academy of Sciences. 115, E4350-E4357. https://doi.org/10.1073/pnas.1718804115.

[18] Urbanek, A.K.; Rymowicz, W.; Mirończuk, A.M. (2018): Degradation of plastics and plastic-degrading bacteria in cold marine habitats. Applied Microbiology and Biotechnology. 102, 7669-7678. doi: 10.1007/s00253-018-9195-y.

[19] Chen, L.; Pelton, R.E.O.; Smith, T.M. (2016): Comparative life cycle assessment of fossil and bio-based polyethylene terephthalate (PET) bottles. Journal of Cleaner Production. 137, 667-676 (2016). https://doi.org/10.1016/j.jclepro.2016.07.094.

Database Development and Special Considerations for Storing Polymer Fate Information

Christopher Rademacher[1], Marina Slijepcevic[2], Tanden Hovey[3], Margaret MacDonell[4]

1. Introduction

The development of plastics has historically focused on obtaining optimal mechanical properties for a product application. This has resulted in materials that are highly desirable for use, but this can come with a heavy environmental burden after that useful life ends. This is because polymers that are developed to be mechanically robust in an application translate to polymers that are very slow to naturally degrade when discarded. As reported by the U.S. Environmental Protection Agency (EPA) for the 2017 year, of 35 million tons of plastic generated, 27 million tons (76%) were landfilled in the United States and 2.9 million tons (8%) were recycled [1]. As landfilled plastics break down over a long period of time, they can leach into the local environment and pose potential harm through various exposure pathways.

A research project has recently begun that aims to identify and synthesize practical information from the scientific literature and other online sources to create an integrated database that provides easy access to practical information about a variety of polymers and their potential fate in the environment following disposal. The aim of the database is to serve as a resource for future polymer developers and producers to help inform their design and manufacture of novel plastics that can be more compatible with the environment after their product life ends.

The potential environmental effects of polymers are broad, varied, and complicated. To help organize the general types of data to be compiled and integrated, basic elements of a simplified conceptual exposure model were outlined to provide a visual overview of environmental fates of waste polymers (Figure 1). The categories in this example concept represent potential sources of releases to the environment, with polymers and associated chemicals then transported to and through various environmental media compartments to human and ecological receptors, for potential exposure through various routes. This simplified model serves as an initial map of the paths a waste polymer might take after its useful life ends. Its components guide the outlining of a data schema, which in turn frames the development of the structure for the database.

[1] Argonne National Laboratory, Lemont, Illinois United States, crademacher@anl.gov
[2] Argonne National Laboratory, Lemont, Illinois United States, mslijepcevic@anl.gov
[3] Argonne National Laboratory, Lemont, Illinois United States, thovey@anl.gov
[4] Argonne National Laboratory, Lemont, Illinois United States, macdonell@anl.gov

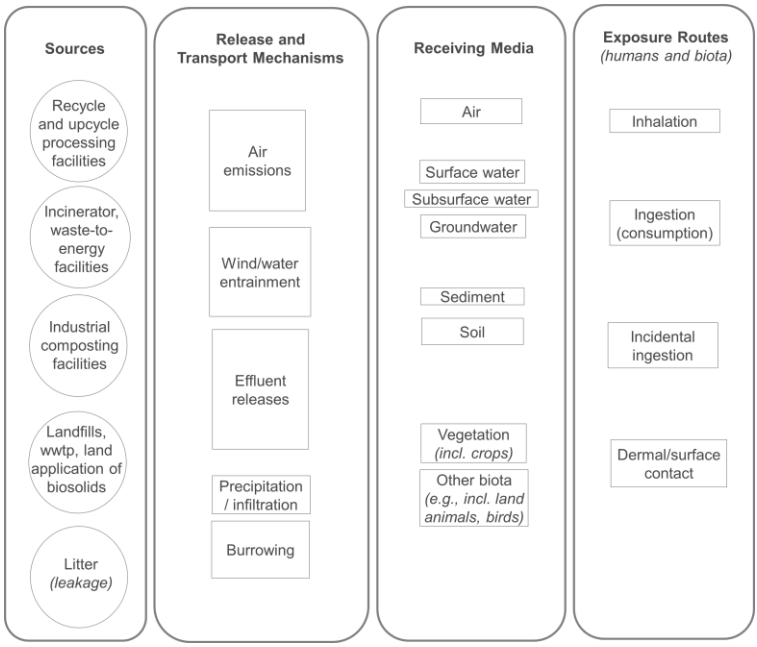


Fig. 1. Simplified conceptual exposure model used to frame the initial data schema.

2. Methods

2.1. Data Gathering

The data gathering step involved scanning the literature for information on each component of the conceptual model, reviewing selected studies to identify data elements relevant to that component, and extracting sample data to help guide the data schema constructed for each topic using the conceptual model as a framework. The initial literature scan involved a screening search conducted in Google Scholar. For some components of the conceptual model, this search produced sufficient data to inform the preliminary database concept. For others, a more systematic approach was employed using the Argonne National Laboratory library search tool, which includes access to Science Direct, Web of Science, patent databases, and more. Forward and backward snowball searches were also conducted, tracing forward to find new literature that cited a given article, as well as backward to find earlier literature cited by that article. Within articles, relevant information is found in tables and figures, as well as within the body of the text.

The Zotero reference management system was used as a library for the candidate literature found, which automatically populates metadata such as author, abstract, and doi number. An ancillary data summary spreadsheet was also created to store other important data contained in the study. In it, data types that are anticipated to be populated as part of developing the prototype database were noted, along with the associated conceptual model component. Other pertinent study metadata not already captured in Zotero

were also recorded, such as a brief article summary along with a description of research needs as stated in the study.

2.2. Automation

The volume of future data is anticipated to be extensive, as research in this area continues to expand, and a considerable amount of information is currently found in tables. For these reasons, it was determined that an automated table extractor would be useful as a first application of automating data extraction for the initial database. However, this presented several challenges, because information in the standard PDF file type cannot easily be located and extracted. Several extractors were implemented and evaluated, including tabula-py, PyPDF2, and camelot. Of these three, the camelot extractor was found to most accurately pull information from the articles because it offers a wide variety of built-in functions that can improve accuracy.

The function that was found to be most useful was one that extracted table information based on a specified location in the PDF document. By drawing a rectangle around the table, camelot more accurately extracted data without extracting extra data. In contrast, automatic table extractions consistently extracted information outside of tables. In order to easily specify the areas in which a table resides, an interactive PDF display was developed using PyMuPDF and Tkinter. In this display (Figure 2a), a box could be drawn around the table, and after saving the table, the next table could be located and selected. This information was then integrated with the data summary document developed during the literature search to identify which table was selected from the given publication. Once the database framework is put in place, the graphical user interface (GUI) enables straightforward and streamlined input of data into the database.

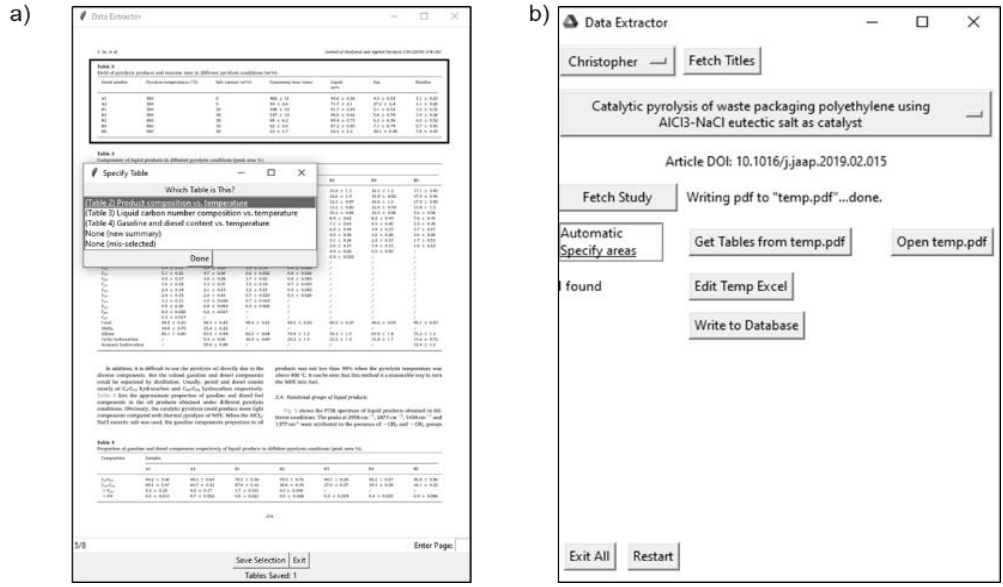


Fig. 2. **a)** PDF reader developed to specify areas where tables are located, for integration with the data summary document to identify the table. **b)** GUI developed to streamline data input.

This PDF reader was implemented in a GUI (Figure 2b) that located the studies found and not yet inputted into the prototype database, and then put them into the database. The interface began with the selection of a study based on those listed in the data summary document. From the selection, the study was located in the Zotero database using pyzotero and was then written to a temporary PDF document using the linked PDF in the Zotero database. Tables could then be either automatically found or found using area specification, which brought up the PDF reader previously described. The tables found were then written to a temporary spreadsheet document, where stored data could be cleaned or modified, if necessary. Finally, the data in the spreadsheet were written to the database.

2.3. Database Development

Before developing a unified data schema, the type of database appropriate for the initial concept was considered. Databases used to store scientific information can be categorized into two primary types: SQL and NoSQL. SQL databases operate relationally, while NoSQL databases operate non-relationally. NoSQL databases have been optimized for the storage of key-value information (and more), which is not true for SQL databases [2]. As part of early planning, a graph structure is being considered for the conceptual data schema, which suggests a NoSQL database.

Several types of NoSQL databases exist, and the type selected for initial testing is a document store database because this type offers versatility and simplicity of implementation. In a document store scheme, items are given a unique key and a document is represented as a value, which can either be structured or unstructured [3]. MongoDB was selected as the database service, which stores documents in a JavaScript Object Notation (JSON)-style format and has straightforward integration with Python, allowing for straightforward integration with other aspects of the project. Information stored in a MongoDB database is readily produced using the MongoDB query syntax. Data temporarily stored in the Excel document can be transferred to the MongoDB database using the Pymongo library.

Because the data being compiled from the literature and other online information sources vary widely, an abstract data document was developed to identify the important features that all data in the database must share. These include the value and units of the datum and associated metadata. As an example, "4.3 tons CO_2 produced per ton plastic burned" would be captured in this format. Because data can exist in a series, the values could exist in an array. The metadata of the datum includes the conditions under which the datum was found, including the value of the independent variable tested, important information about how the data were found, and a citation to the study from which the datum was extracted. This study citation is itself another document, which includes some essential information for identifying the study, including the doi number, title, brief summary, and authors. This structure is demonstrated in Figure 3. Not all data have this general structure, and fields left blank are automatically handled by MongoDB.

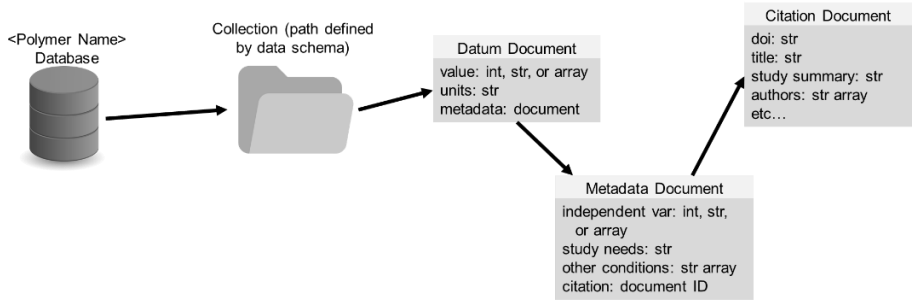


Fig. 3. Structure of the database concept and general structure for the document corresponding to a single datum.

3. Discussion

Nearly 400 candidate studies were compiled in Zotero through July 2020, and about half were evaluated in detail. From those evaluations, 362 items were identified as possible sources of data to incorporate in the upcoming database; 129 of these have been captured in a data summary document, covering nearly all components of the conceptual exposure model. Information contained in these studies originates from scientific publications, government reports, books, existing databases, and various other sources. Each source was screened to assess the reliability of the data. Because broad awareness of the waste plastics in the environment is increasing, it is anticipated that the number of publications will likewise increase as further research is funded to address gaps in our current understanding of the environmental fate of these materials and associated risks. Further data will provide a foundation for machine learning applications in a future stage of this work.

A portion of the data schema derived from the conceptual exposure model is presented in Figure 4, showing the types of data associated with pyrolysis and incineration processes. This schema illustrates data relevant to the conceptual model components of incinerators and recycle/upcycle processing facilities. Data schemas were also developed for the other components of the conceptual model.

The prototype data schema developed from the literature reviewed to date is being used to structure a preliminary MongoDB database, and the GUI developed to streamline data inputs was used in populating the information. As illustrative data were input from online sources, the accuracy of the extracted data was monitored by reading the extracted data into a temporary spreadsheet document. In some cases, specifying the table areas resulted in no table being found due to an issue in selecting the entire table area. This was addressed in the code by allowing for reselection of the table before saving. Additionally, in some cases, the code extracted information from the text surrounding the tables, which necessitated manual cleanup before writing to the database. Because the amount of literature being written to the database is currently small enough, no automatic implementation of this cleanup has been needed to date (thus, none has been written to date).

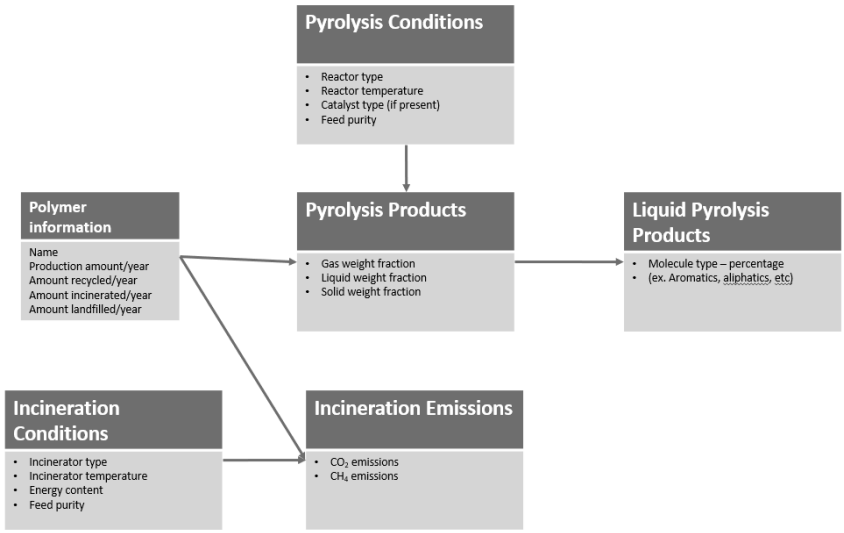


Fig. 4. Example data schema for pyrolysis and incineration categories in the conceptual model.

4. Conclusion

Diverse data on polymer properties can be found online that are relevant to their fate in the environment after the end of a product's useful life. These data exist in a variety of sources, and an approach has been developed to streamline their compilation to support the development of a prototype database. The overall aim is to create a practical database that combines information about the characteristics of polymers with data on the environmental fate of plastics through several disposition and exposure paths. When completed, this database is envisioned to inform the design of new polymers that are more biofriendly, by considering their potential post-use fate environmental from the outset.

Online publications can be readily organized into a Zotero database with data summaries stored in an ancillary spreadsheet. To simplify writing the information from data tables to a MongoDB database, the camelot PDF table extractor was found to be an effective tool, with implementation via a GUI that was built in Python. Further development includes supporting the extraction of information from other online sources, such as from text contained in PDF documents and direct scraping from the Internet. This preliminary database has advanced planning for a practical integrated resource anticipated to be available for internal testing in 2022.

Acknowledgements

The authors gratefully acknowledge Cristina Negri, Director of the Environmental Science Division at Argonne National Laboratory, for her vision in developing and implementing this research project. We also thank Young-Soo Chang for his technical review of this work. We further gratefully acknowledge Alicia

Lindauer, Andrea Bailey, Nichole Fitzgerald, and additional colleagues in the U.S. Department of Energy, Office of Science, Bioenergy Technologies Office, for their support and funding for this research. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. http://energy.gov/downloads/doe-public-accessplan.

References

[1] U.S. Environmental Protection Agency. (2019): Advancing Sustainable Materials Management: 2016 and 2017 Tables and Figures.

[2] Li, Y.; Manoharan, S (2013).: A performance comparison of SQL and NoSQL databases. In: 2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM). pp. 15-19. IEEE, Victoria, BC, Canada.

[3] Sharma, V.; Dave, M. (2012): SQL and NoSQL Databases. IJARCSSE. 2, 20–27.

Developing a Preliminary Data Structure to Assess Plastics in Freshwater Environments

Sneha Nachimuthu[1], Jennifer Cronin[2], Margaret MacDonell[3]

1. Introduction

A significant fraction of all plastics ever produced remains in our lands and waters because these products were explicitly designed to resist degradation. Decades of sustained increases in the production of plastics have resulted in the global accumulation of waste plastics in the environment. To help mitigate this problem, an integrated understanding of the environmental fate of plastics is needed, to help inform the design of replacement polymers that are more environmentally benign.

This study focuses on the fate of plastics in freshwater systems. Environmental plastics can be split into two groups based on size: macroplastics and microplastics. Although the dimensions used in the literature to distinguish between these two categories can vary, this assessment applies those most commonly used. That is, microplastics are smaller than 5 millimeters (mm) and are often further classified into types of plastics like microfibers, fragments, microbeads, films, and foams [2]. Macroplastics are at least 5 mm and can take many different forms. Macroplastics can also be broken down or degraded into microplastics as they are transported through the environment via various mechanisms and deposited in different environmental media. Some researchers identify an additional category, mesoplastics, with dimensions between 2.5 and 5 mm [11]; however, specific data in this size range are limited and there is no clear benefit to including this additional group. For this reason, the current assessment implements the common definitions unless otherwise indicated (e.g., per definitions used in a particular study).

In a freshwater setting, multiple types of environmental media are present, and each can have different properties that affect the way plastics behave. Microplastics have been found in many lakes and sediments, and they can originate from urban and other terrestrial sources; those found in rivers are a potential source of plastics entering the oceans [10, 13].

Microplastics can also be transported to soils through various routes, including applications of soil amendments, lake and river flooding, atmospheric deposition, and biotic transport. Soils are sinks for microplastics because once they enter the soil environment they can persist for many decades; some estimate it could take many hundreds of years to fully degrade into their basic components [3]. For this reason, as more microplastics are added to soil, such as through biosolids, they will continue to accumulate there [7]. In certain soil types and under certain environmental conditions, microplastics can also migrate downward through soil to groundwater [9].

Microplastics have been detected in plastic water bottles, and at varying levels in specific locations where drinking water is sourced [8]. Microplastics have been found in a karst system that serves as the

[1] Argonne National Laboratory, Lemont, Illinois United States, snachimuthu@anl.gov
[2] Argonne National Laboratory, Lemont, Illinois United States, jcronin@anl.gov
[3] Argonne National Laboratory, Lemont, Illinois United States, macdonell@anl.gov

source of drinking water for the region [12]; they have also been detected in other freshwater sources that serve as a drinking water sources. Additives and other chemicals associated with these microplastics pose a potential concern for human health, depending on the magnitude and duration of exposures. Microplastics can also provide an effective surface for the sorption of other chemical pollutants from the local environment, thus introducing a further potential for exposure. Such pollutants have been found adsorbed to microplastics in both marine and coastal areas [1].

Some studies have indicated the potential for adverse ecological health effects when plastics are introduced to earthworms, bivalves, and other organisms. Impacts to tissues and the immune system have been reported for earthworms, while intestinal and oxidative damage has been reported in nematodes [9]. With the potential to accumulate through trophic levels, microplastics have been found in consumer goods ranging from seafood and salt to honey and beer [9].

The source of a given plastic's introduction into the environment may be indicative of its type. Regarding microplastics, certain types have been associated with particular sources; as an example, wastewater treatment plants (wwtps) can be a major source of microfibers in the environment, originating from plastic microfibers in clothes that are agitated in washing machines. Because these treatment facilities are not equipped with technology capable of filtering out plastic microfibers, these are simply released into the environment [15]. Cosmetics like glitter and microbeads found in exfoliants are also among the many sources of microplastics released from these facilities [6]. Sludges or biosolids containing microplastics are often found in runoff and can result from fertilizers and plastic mulches used to help reduce soil erosion and lessen pesticide use [14]. Packaging is commonly made of Styrofoam or polystyrene (PS), which is very susceptible to degradation into microplastics and is also more easily transported to marine environments [4]. If unidentified plastics can be analyzed, certain additives associated with specific sources or types could potentially enable forensic tracking through different environmental media.

Attention to waste plastics in freshwater ecosystems is relatively new, with a shift from the dominant focus on marine and coastal systems occurring within the last several years. For this reason, related information and data analysis techniques are still evolving, and overall data syntheses are still being developed. A main objective of the current literature evaluation is to identify key data elements that can underpin a practical structure for the database concept being developed. The aim is to outline a structure that can effectively incorporate the limited information available from existing studies while maintaining flexibility to evolve as other data become available to fill the substantial gaps in current knowledge.

2. Methods

A literature search was conducted from May through July 2020 through the Argonne National Laboratory research library system, which accesses several databases including ScienceDirect, and Web of Science. Of forty candidate papers identified from the peer-reviewed literature, fifteen were examined in detail. These papers were relevant to both microplastics and macroplastics in riverine, lake, and groundwater environments. Soil and sediment linked to freshwater were also included.

3. Results

Data found in the literature related to plastics in freshwater systems reflect discrepancies due to differences in data collection and reporting methods. It appears that recent studies are beginning to follow a more consistent methodology and/or ascertaining what methods provide the most accurate results. Information from earlier studies must be carefully extracted and recorded to qualify the data that will be incorporated into the database that is being developed to inform the design of alternate polymers that are more environmentally compatible. The data for freshwater settings are far from uniform across the various studies.

For example, metrics commonly reported to quantify plastics in the environment are particles per area or particles per volume. However, these metrics do not readily lend themselves to integrating data across different studies because the particle density is not reported, so the mass cannot be identified. As a result, it is difficult to compare or combine these disparate data to develop a fuller understanding of the levels of macroplastics and microplastics in freshwater systems. Nevertheless, these data do provide some insight into where further attention is warranted to better understand the potential impacts of plastics in certain environments. When extracting and synthesizing data from the literature, it will be important to include the metadata that provide important information about the sampling and analysis methods to support an evaluation of the quality of these data (see Table 1; note FT-IR = Fourier-transform infrared method).

Medium	Plastic Size or Shape	Analysis Method	Unit of Measure	Abundance	Location	Reference
Surface water	Microplastic	FT-IR; Raman	p/m^3	Range: 0.05 - 32, median: 1.9	Great Lakes, U.S.A.	Baldwin [2]
Surface water	Macroplastic	Visual	items, g, m^2, L, m	217.5 items, 1232.8 g	Setubal Lake, Portugal	Blettler [4]
	Mesoplastic	Visual		25 items, 1.9439 g		
	Microplastic	FT-IR		704 items, 0.074 g		
Surface water	Microplastic	Mass-Spec	p/m^3	Range: 0.9±0.4 - 13±5	Ofanto River, Italy	Campanale [5]
Ground water	Microplastic	Dissecting microscope Mass-Spec	p/L	Range: 15.2, median: 6.4	Illinois Karsts, U.S.A.	Panno [12]
Sediment	Microbead	Dissecting microscope	p/m^2	Median: 52, mean: 13759±13685	St. Lawrence River, Canada	Castañeda [6]
Sediment	Microplastic	μ-FT-IR	kg^{-1} dry weight	Mean: 80.2±59.5	Shanghai, China	Peng [13]
Soil	Microplastic	Stereo microscope	p/5g		Chile	Corradini [7]
Soil	Microplastic	Stereo microscope, μ-FT-IR	items/kg	Shallow soil abundance, 78±12.91; deep soil abundance, 62.5±12.97	Shanghai, China	Liu [11]
	Mesoplastic			Shallow and deep soil abundance, 3.25±1.04		

Table 1. Highlights of diverse studies on plastics in different environmental settings.

Differences in plastic types and sizes, environmental settings, and transport mechanisms must be considered when outlining a data schema for the database being developed. It is useful to distinguish between the environmental setting in which a plastic is found and the impact that plastic may have on those media and the organisms that could be exposed. Because plastic particles can be readily transported from one location to another in certain settings, the database design must consider how to reflect the fate and transport of these particles over time. The potential impact of these plastics may change in different settings, which makes this aspect harder to quantify. It will be important to characterize certain parameters specific to each environmental medium, to establish the best understanding of factors affecting the behavior of a given plastic or microplastic mixtures in various environments (Figure 1).

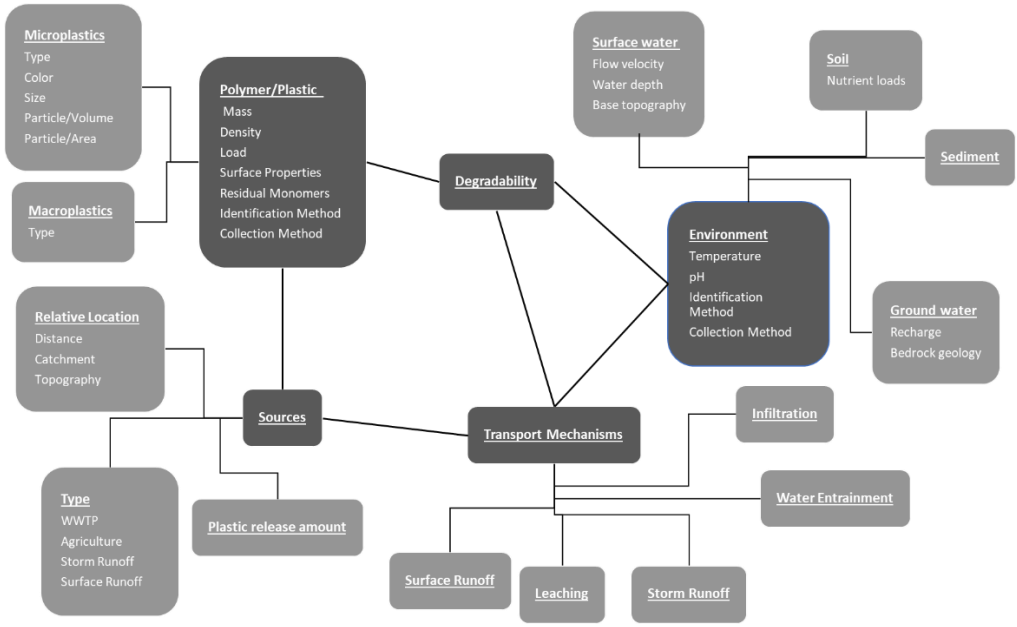


Fig. 1. Illustrative data schema for plastics in freshwater systems and associated media.

4. Discussion

Public datasets available from a variety of organizations serve as valuable sources of information for the integrated database being envisioned to enable data-driven consideration of post-use environmental implications from the outset, during the design of a new polymer. Geospatial aspects offer the glue for combining disparate datasets, as illustrated by the Spatial Data Infrastructure (SDI) community related to data modeling and geospatial data sharing via standardized web services. This work emphasizes and enables interoperability with complementary datasets and geoanalytical tools. The concept of managing plastic-related data as a layer in a geographic information system (GIS) is fundamental to integrating data on microplastics with other well-known geospatial datasets (including topography, soil types, and more) and leveraging spatial analyses that examine correlations across different layers of information in the same area of interest. Thus, going forward, as more data become available to characterize plastics across different

environmental settings, one way to effectively integrate these data is to organize them geospatially so that possible sources and transport mechanisms can be more easily identified (Figure 2).

Levels that might be incorporated include the following considerations:

- Overall topography, soil types, and location of a watershed are important to understanding the flow of water in an area and can inform where plastics might settle and concentrate, using other variables. Data could potentially be extracted from existing sensor systems for parameters such as velocity and discharge of water (e.g., from the U.S. Geological Survey water sensors within the United States).
- Land use might be correlated with certain types of plastics; it will be important to also include context for other variables.
- Precipitation could increase particle transport and can affect where certain plastics are carried in runoff.
- Seasonal patterns can also affect plastics and degradation [11] and should be taken into account.
- Location of release sources will affect overall distribution, so identifying sources of microplastics and macroplastics can support the tracing of transport pathways.
- Studies that pinpoint the geographic areas and media investigated can help develop a fuller picture of plastic release and transport through different environmental compartments.

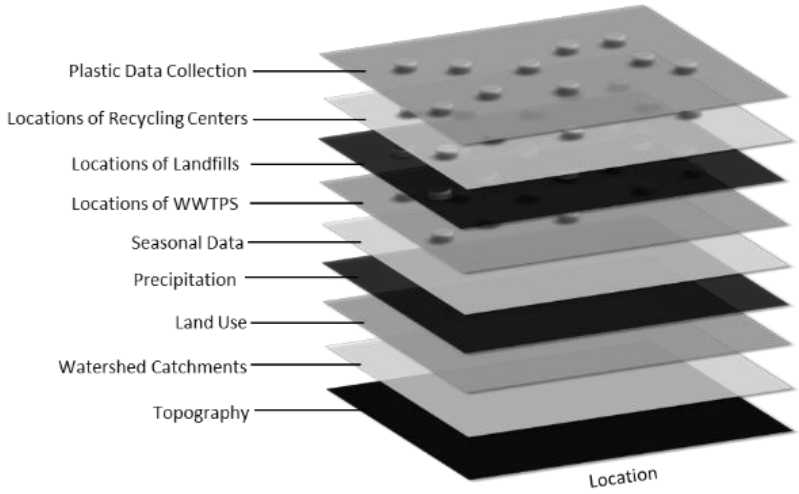


Fig. 2. Example geospatial layers to consider when organizing data on environmental plastics.

Identifying the locations and amounts of plastics that have been transported through various environmental compartments can also inform future choices regarding what types of plastics and additives could be more or less suitable for use (and potential release) in different settings. The function and durability of the plastic are also important considerations.

A better understanding of the nature and extent of microplastics in freshwater environments can help inform the development of novel plastics, as well as strategies to minimize environmental impacts (e.g., related to the placement of wastewater treatment plants or use of mulch containing plastics). The next phase

of this project will focus on the informatics aspects of the database development work. Looking ahead, it will be important for data collection and analysis methods for microplastics and macroplastics in different environmental settings to be standardized, to enable the integration of limited data that can help frame the design of new plastics to be more biobenign.

Acknowledgements

The authors gratefully acknowledge Cristina Negri, Director of the Environmental Science Division at Argonne National Laboratory, for her vision in developing and implementing this research project. We also thank Young-Soo Chang for his technical review of this work. We further gratefully acknowledge Alicia Lindauer, Andrea Bailey, Nichole Fitzgerald, and additional colleagues in the U.S. Department of Energy, Office of Science, Bioenergy Technologies Office, for their support and funding for this research. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program. The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. http://energy.gov/downloads/doe-public-accessplan.

References

[1] Akdogan, Z.; Guven, B. (2019): Microplastics in the environment: A critical review of current understanding and identification of future research needs. Environmental Pollution. 254, 113011. https://doi.org/10.1016/j.envpol.2019.113011.

[2] Baldwin, A.K.; Corsi, S.R.; Mason, S.A. (2016): Plastic debris in 29 Great Lakes tributaries: Relations to watershed attributes and hydrology. Environmental Science Technology. 50, 10377–10385. https://doi.org/10.1021/acs.est.6b02917.

[3] Bläsing, M.; Amelung, W. (2018): Plastics in soil: Analytical methods and possible sources. Science of The Total Environment. 612, 422–435. https://doi.org/10.1016/j.scitotenv.2017.08.086.

[4] Blettler, M.C.M.; Ulla, M.A.; Rabuffetti, A.P.; Garello, N. (2017): Plastic pollution in freshwater ecosystems: macro-, meso-, and microplastic debris in a floodplain lake. Environmental Monitoring Assessment. 189, 581. https://doi.org/10.1007/s10661-017-6305-8.

[5] Campanale, C.; Stock, F.; Massarelli, C.; Kochleus, C.; Bagnuolo, G.; Reifferscheid, G.; Uricchio, V.F. (2020): Microplastics and their possible sources: The example of Ofanto river in southeast Italy. Environmental Pollution. 258, 113284. https://doi.org/10.1016/j.envpol.2019.113284.

[6] Castañeda, R.; Avlijas, S.; Simard, M.A.; Ricciardi, A. (2014): Microplastic pollution in St. Lawrence River sediments. Canadian Journal of Fisheries and Aquatic Sciences. 71, 1767–1771. https://doi.org/10.1139/cjfas-2014-0281.

[7] Corradini, F.; Meza, P.; Eguiluz, R.; Casado, F.; Huerta-Lwanga, E.; Geissen, V. (2019): Evidence of microplastic accumulation in agricultural soils from sewage sludge disposal. Science of The Total Environment. 671, 411–420. https://doi.org/10.1016/j.scitotenv.2019.03.368.

[8] Eerkes-Medrano, D.; Leslie, H.A.; Quinn, B. (2019): Microplastics in drinking water: A review and assessment. Current Opinion in Environmental Science & Health. 7, 69–75. https://doi.org/10.1016/j.coesh.2018.12.001.

[9] He, D.; Luo, Y.; Lu, S.; Liu, M.; Song, Y.; Lei, L. (2018): Microplastics in soils: Analytical methods, pollution characteristics and ecological risks. TrAC Trends in Analytical Chemistry. 109, 163–172. https://doi.org/10.1016/j.trac.2018.10.006.

[10] Lebreton, L.C.M.; van der Zwet, J.; Damsteeg, J.-W.; Slat, B.; Andrady, A.; Reisser, J. (2017): River plastic emissions to the world's oceans. Nature Communications. 8, 15611. https://doi.org/10.1038/ncomms15611.

[11] Liu, T.; Zhao, Y.; Zhu, M.; Liang, J.; Zheng, S.; Sun, X. (2020): Seasonal variation of micro- and meso-plastics in the seawater of Jiaozhou Bay, the Yellow Sea. Marine Pollution Bulletin. 152, 110922. https://doi.org/10.1016/j.marpolbul.2020.110922.

[12] Panno, S.V.; Kelly, W.R.; Scott, J.; Zheng, W.; McNeish, R.E.; Holm, N.; Hoellein, T.J.; Baranski, E.L. (2019): Microplastic Contamination in Karst Groundwater Systems. Groundwater. 57, 189–196. https://doi.org/10.1111/gwat.12862.

[13] Peng, G.; Xu, P.; Zhu, B.; Bai, M.; Li, D. (2018): Microplastics in freshwater river sediments in Shanghai, China: A case study of risk assessment in mega-cities. Environmental Pollution. 234, 448–456. https://doi.org/10.1016/j.envpol.2017.11.034.

[14] Qi, R.; Jones, D.L.; Li, Z.; Liu, Q.; Yan, C. (2020): Behavior of microplastics and plastic film residues in the soil environment: A critical review. Science of The Total Environment. 703, 134722. https://doi.org/10.1016/j.scitotenv.2019.134722.

[15] Re, V. (2019): Shedding light on the invisible: addressing the potential for groundwater contamination by plastic microfibers. Hydrogeology Journal. 27, 2719–2727. https://doi.org/10.1007/s10040-019-01998-x.

A Database on the Health Risks of Plastics

Marina Slijepcevic[1], L'Nazia Edwards[2], Aijalon Kilpatrick[3], Phuong Khanh Tran Nguyen[4], Margaret MacDonell[5]

1. Introduction

The production, use and post-use management of plastics can all involve multiple chemical processes and releases. Furthermore, for used plastics that are accumulating in the environment, degradation products and the potential for impacts to ecological and human health over the long term are poorly understood. Quantitative information relevant to the end-of-life fate of plastics is generally limited, variable, and scattered across multiple sources. Therefore, a centralized database that provides ready access to the different types of data needed to inform planning for new plastics with a more benign footprint is needed, with an ultimate goal of advancing the circular economy.

A project to develop such a consolidated database has begun. The aim is to provide a practical resource that can be used to inform the design of new polymers with much lower end-of-life impacts than conventional plastics, by considering environmental implications from the outset. Bioplastics are a particular emphasis. The database will be populated with data in the public domain, and when completed (within three years), the plan is to make it available online for broad access.

The first step of this work focuses on assessing the state of knowledge to identify data elements that address characteristics of polymers and associated additives, characteristics of environmental settings that affect the fate of these materials, and factors that affect human and ecological exposures and effects. Characterizing potential exposure pathways is a key aspect of environmental health risk analyses. For environmental plastics, core elements of an exposure assessment include: (1) the source(s) and compositions, including associated pollutants; (2) release mechanisms by which these materials enter the environment; (3) the environmental media that receive these materials (e.g., soil, water, air); and (4) the routes by which humans and other biota can be exposed (e.g., ingestion, inhalation, and absorption across contact surfaces). Results of the exposure assessment is combined with toxicity data to estimate risks.

The literature evaluation conducted in the first phase of this project is anticipated to uncover specific knowledge gaps that could be used to help inform research planning, from agency portfolios to program initiatives in academia and other research organizations. In developing the database, promoting interoperability will be a central aspect, to enable those in the private sector (e.g., the plastics industry) to link with proprietary data or software, to further support internal efforts toward a circular economy.

[1] Argonne National Laboratory, Lemont, IL, USA, mslijepcevic@anl.gov
[2] Argonne National Laboratory, Lemont, IL, USA, ledwards@anl.gov
[3] Argonne National Laboratory, Lemont, IL, USA, akilpatrick@anl.gov
[4] Argonne National Laboratory, Lemont, IL, USA, ptrannguyen@anl.gov
[5] Argonne National Laboratory, Lemont, IL, USA, macdonell@anl.gov

2. Methods

2.1. Literature search

A preliminary literature search was conducted using multiple search engines through Argonne National Laboratory's library system. Information sources included the scientific literature as well as agency databases and other online data. Keywords included: plastic, polymer, additive, plasticizer, plastic manufacturing, byproduct, contaminant, environment, human health, ecological, hazard, harm, toxic/toxicity, brominated flame retardant (BFR), polyvinyl chloride (PVC), phthalate. For an exploratory search, candidates were selected for plastic type and environmental transport setting -- notably PVC and BFR, and aquatic environment. The types of data identified from this literature search will serve as the basis for the initial data schema.

2.2. Data compilation

Candidate publications were downloaded into the Zotero reference management system, and a shared data summary Excel file was created to capture descriptions of the types of data found. Data extraction was implemented with Python programming. Webpage and pdf data extractions were implemented to scrape data from selected tables. The use of webpage html data extraction approaches raised an issue related to scraping constraints. More specifically, certain websites (including ScienceDirect and OnlineWiley) require special access keys referred to as Application Programming Interface (API) keys for each web address, or Uniform Resource Locator (url). Extensible Markup Language (XML) versions of scientific article pages were accessible with Elsevier's API key service, as detailed in its text mining documentation.

Several scientific databases were reviewed to select an appropriate platform for an initial prototype database concept. This analysis found that government databases commonly utilize MySQL frameworks. Meanwhile, the MongoDB platform was identified as a good fit for the preliminary database concept for this project. This platform was selected in part because it stores JavaScript Object Notation (JSON) files, which are suitable for large amounts of data. MongoDB can also accommodate variable data structures and can implement a hierarchical database; the latter is desirable because it is consistent with the shape of the data schema being developed for the project, which will capture multiple properties of plastics and environmental settings comparison.

3. Discussion

3.1. Chemical toxicity

Several adverse health effects have been associated with polymers and residuals across various stages of the plastic life cycle. For example, manufacturing related to both virgin and recycled plastics could affect human health and the environment through associated energy consumption and emissions or releases of byproducts. Some studies report effects of microplastics on the reproductive systems of zooplankton, while human toxicokinetics data indicate that microplastics can potentially pass through the gut after eating food that contains them [5, 11]. Toxicity values used to assess human health effects from chronic (lifetime)

exposures can be found for selected chemical components of plastics or their environmental degradation products in the U.S. Environmental Protection Agency (EPA) Integrated Risk Information System (IRIS) [12]. Furthermore, human and ecological toxicity values are available through the European Chemicals Agency (ECHA) program for specific polymers managed under the European Union (EU) program for Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH). However, standard toxicity estimators are unavailable for most formulations for plastics, because such chemicals are typically covered by confidential business information and other proprietary data protection measures.

Toxicity concerns are illustrated by an example plastic, PVC, which is the world's third most widely produced synthetic polymer and can be one of the most dangerous when burned in urban, open-air settings. This heat-resistant polymer has been manufactured with chemicals such as phthalates, bisphenol A (BPA), lead, dioxin, and cadmium, which can pose health concerns at elevated exposure levels [1]. Consumer products (including building materials, clothing, medical devices, children's toys, and food packaging) contain PVC additives, including phthalates, and the list of products continues to grow. Exposures to these additives include ingestion via food, enteral nutrition formulas, and nutritional supplements, as well as pharmaceuticals, and mouthing of toys and other objects (by young children) [1]. Phthalates make PVC more flexible and harder to break, and they include di-(2-ethylhexyl) phthalate (DEHP), di-isononyl phthalate (DiNP), and di-isodecyl phthalate (DiDP) [2]. These additives can leach from products because they are not chemically bound to the plastic matrix, and they are generally lipophilic, which affects their partitioning [3]. The phthalate DEHP is reported as the most commonly inhaled; this chemical is often found in paint, newly furnished houses, and PVC manufacturing and handling facilities [4]. Prenatal exposure to phthalates have been linked with neurodevelopmental, reproductive, and immune system effects, with potential implications for future obesity as well [5].

Flame retardants represent another example of additives commonly used in polymers. These chemicals reduce the chance of a fire starting and/or reduce the severity of a fire in a wide assortment of plastic products [6]. Over the past two decades, the use of certain additives referred to as legacy BFRs has been reduced because of potential health concerns. These legacy compounds include polybrominated diphenyl ethers (PBDEs), hexabromocyclododecane (HBCD), and tetrabromobisphenol A (TBBPA). These additives have been linked to endocrine, reproductive, and behavioral effects in laboratory animals; they can also reduce T4 levels (a thyroid hormone), which could lead to impaired brain growth in animals and humans if abnormal levels are sustained over time. Toxicity studies have also indicated reproductive system effects. For example, the circulation of estradiol in rats is reduced when they are exposed to a common PBDE congener, BDE-47. Delayed puberty, reduced testosterone, reduced sperm count, and reduced ovary weight have also been linked to exposures to common PBDE congeners. A study of male pups exposed to TBBPA reported an increase in testes weights, testosterone levels, and aromatase activity. A similar study with HBCD reported developmental effects in exposed offspring related to decreased trabecular bone mineral density. Additionally, PBDEs have affected locomotion activity and induced cognitive impairment in mice and rats, while HBCD and TBBPA affected dopamine-dependent activity and hearing function [7]. The data found in these studies are being used to outline specific data elements for the database.

Information related to policies, directives, or other administrative controls is also being compiled. For example, legacy BFRs such as penta- and deca-BDE were added to the list of persistent organic pollutants (POPs) under the Stockholm Convention in 2009, and HBCD was added in 2013. Although the European

Union (EU) reported in 2006 that TBBPA is unlikely to cause health risks, a 2017 study suggested that it be classified as an endocrine disruptor for mammals [7]. Compared with the legacy BFRs, much less toxicological information exists for "novel" BFRs (NBFRs). These substances are only recently entering the environment, and they are not considered to pose a similar health concern as their predecessors [7]; the potential for human exposures and effects over the long term is not yet well understood.

The toxicity data for legacy BFRs and other additives identified during the initial literature review are being used to outline data elements for the database. Example elements for categorizing additives and organizing selected toxicity information are presented in Table 1; BDE-47 is used to illustrate part a.

Additive Type	**Additive Class**	**Additive Name**	**Associated Polymers**
Example: Broad additive family	*Example: Classification within additive family*	*Example: Name (Abbreviation)*	*Example: Polymers with this constituent*
Flame retardant	BFR	2,2',4,4'-Tetrabromodiphenyl ether (BDE-47)	Acrylonitrile butadiene styrene (ABS), Polypropylene (PP), Polycarbonate (PC), Polystyrene (PS), Epoxy

Table 1a. Data elements related to classification of additives.

Exposure Route	**Organ/System**	**Effect Level**	
Example: Inhalation, ingestion, dermal absorption	*Example: Thyroid/endocrine system*	*Example: No observed adverse effect level (NOAEL)*	*Example: Lowest observed adverse effect level (LOAEL)*

Table 1b. Data elements related to human toxicity of additives.

Toxicity studies span multiple animal species, while few data exist for human exposures and effects. Thus, context for extrapolations commonly applied to inform human toxicity estimators are also considered in outlining the data elements, including effect thresholds (e.g., NOAEL and LOAEL). An example set of data elements related to additives is presented in Figure 1.

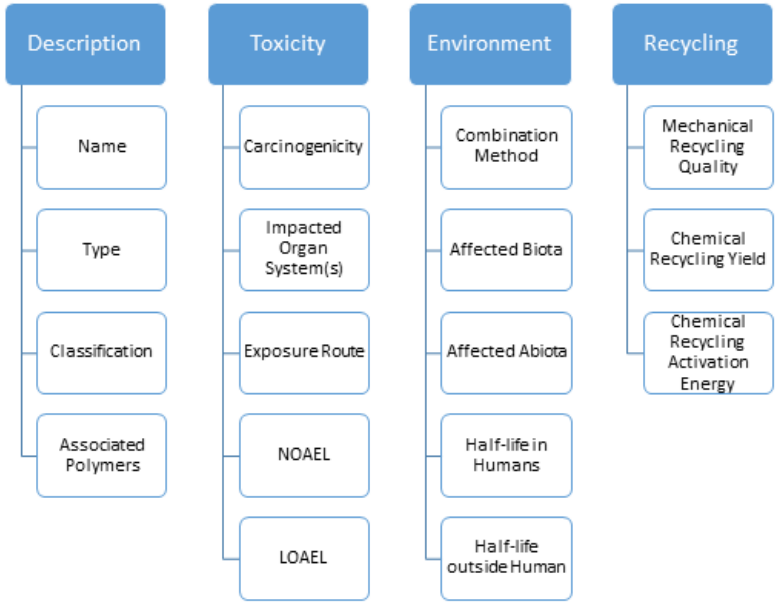


Fig 1. Example data categories relevant to additives derived from the initial publication set.

3.2. Transport and Fate

The transport and fate of plastics in the environment influence the potential for exposures and effects. Taking an aquatic setting as an example, additives can leach from plastic debris into the surrounding water, while contaminants already in the water can sorb onto this debris. Polyethylene, polypropylene, and polystyrene have been reported to sorb organic compounds from natural surface waters (e.g., humic substances), while polyethylene terephthalate (PET) and PVC also sorb these compounds but at much lower levels [9]. The sorption potential is affected by surface area (among other things, including the presence of biofilm), and surface area is affected by particle size; thus, both constitute data elements to be included in the database. Plastic composition and structure will also affect degradation rates, as do environmental conditions. For this reason, it will be important to link polymer and setting characteristics with the degradation process type – e.g., biodegradation, photo-degradation, thermo-oxidative degradation, and hydrolysis [10].

Recent studies on microplastics and even smaller particles indicate that several factors affect their bioavailability [11]. The effect of trophic transfer through the food chain is not well understood, but this transfer has been reported in marine and freshwater systems [6]. In short, the initial literature review has revealed an array of data elements to consider in developing the schema for the database being developed. These categories span physicochemical properties; particle size; sorption potential; and degradation mechanisms, rates, and fate products. Environmental conditions that influence degradation include pH and UV radiation. Data gaps identified within these publications are also being compiled, for incorporation into

the database to help inform research planning. Figure 2 illustrates example data elements related to environmental transport, fate, and exposures.

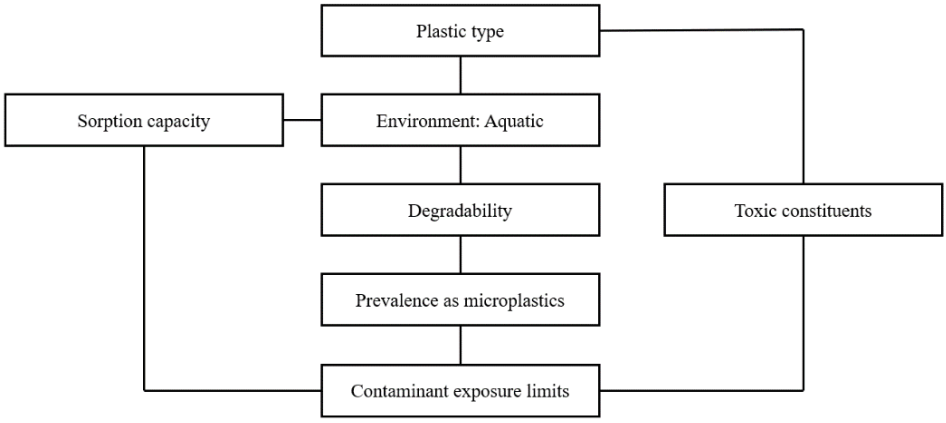


Fig. 2. Example data elements related to environmental exposures.

3.3. Additional data sources

Other information resources examined include the EPA Toxics Release Inventory (TRI), which tracks chemicals considered "carcinogenic or can cause other chronic health effects, could produce significant adverse acute health effects, and have significant adverse environmental effects" [13]. TRI requires annual reports from any manufacturing facility whose operations use, produce or process designated chemicals above a threshold level. The TRI database includes the plastic and rubber industry sector, and facilities are categorized by the type of product, such as tire, pipe fitting, and plastic film. Because few studies were found to compare emissions from production of different types of plastics, the TRI categories are helpful in offering context for emission profiles for several major types of plastics.

Although the TRI data extend across the United States and cover a variety of industries, they do not provide sufficient granularity for quantitative analyses at the level of interest for this project. For example, the "plastic bags and pouch" category might include high density polyethylene (HDPE), low density polyethylene (LDPE), linear LDPE (LLDPE), and potentially other bio-based plastics; furthermore, the proportion of each type is unknown. A different dataset that can provide more granularity for certain aspects is the Franklin Associate's cradle-to-gate emission and energy use inventory for nine types of plastic resins [14]. However, this dataset lacks emission profiles for different industries and does not detail potential adverse effects associated with any emissions, because the scope is limited to an inventory. Insights from considering each of these sources are helpful in framing the plan for an integrated database that aims to serve as a useful resource for different types of users and applications**.**

Reports by organizations such as The Center for International Environmental Law (CIEL) and International Energy Agency (IEA) also serve as valuable information sources, including for the references cited therein. Examples include the plastic industry's global carbon budget, and best and worst case scenarios for energy consumption or resource depletion in the context of global policies or goals. The data

identified in this initial literature search regarding local wastewater treatment plants (WWTPs) and plastic manufacturers indicate that plastic production sites constitute a major source of microplastic pollution, yet they are typically understudied and frequently unaccounted for [15].

Reporting challenges include those related to mixed-waste streams. Meanwhile, laboratory data and case studies can provide useful details, but their scopes are typically limited and extrapolations are difficult to support. Many researchers draw from inventories to conduct life cycle analyses (LCAs) or risk assessments, and some have explored assigning risk categories to monomers and additives based on existing international harmonized systems and inventories [16]. Such studies provide valuable context for the planned database, to assure that it will capture data elements important to LCAs and risk analyses.

4. Conclusion

The scientific literature and other online information sources investigated during the first stage of this project provide useful context regarding the state of knowledge for plastics in the environment, from disposition and fate to potential exposures and effects. Quantitative data relevant to the environmental fate of polymers and associated chemicals are limited, varied and scattered. Studies of specific plastics cannot be extrapolated or generalized to other polymers or environmental conditions. Nevertheless, the targeted consideration of selected example plastics and additives such as PVC and flame retardants has been valuable for defining a number of data elements that will anchor the data schema and ultimately the database for this project. In addition, the toxicity data uncovered in the initial literature search and synthesis are being used to extend the conceptual exposure model for plastics in the environment to encompass dose-response and risk data. A key driver is the need for readily accessible information integrated across multiple technical themes to better inform the innovation of polymers that are more environmentally responsible. The objective is to create a practical database that enables those interested in developing more biobenign polymers to consider end-of-life environmental implications at the outset, during the design phase.

Acknowledgements

The authors gratefully acknowledge Cristina Negri, Director of the Environmental Science Division at Argonne National Laboratory, for her vision in developing and implementing this research project. We also thank Young-Soo Chang for his technical review of this work. We further gratefully acknowledge Alicia Lindauer, Andrea Bailey, Nichole Fitzgerald, and additional colleagues in the U.S. Department of Energy, Office of Science, Bioenergy Technologies Office, for their support and funding for this research. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship (SULI) program, and by the Office of Environmental Management under the Minority Serving Institutions Partnership Program (MSIPP). The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and

display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. http://energy.gov/downloads/doe-public-accessplan.

References

[15] Schettler, T. et al. (2006): Human exposure to phthalates via consumer products. In: International Journal of Andrology 29, 134-139.

[16] Proshad, R.; Kormoker, T.; Islam, M. S.; Haque, M. A., Rahman, M. M.; Mithu, M. M. (2017): Toxic effects of plastic on human health and environment : A consequences of health risk assessment in Bangladesh. In: International Journal of Health 6(1), 1-5.

[17] Yong, C.Q.Y.; Valiyaveetill, S.; Tang, B. L. (2020): Toxicity of microplastics and nanoplastics in mammalian systems. In: International Journal of Environmental Research and Public Health 17(5), 1509.

[18] Miao, Y.; Wang, R.; Lu, C.; Zhao, J.; Deng, Q. (2016): Lifetime cancer risk assessment for inhalation exposure to di(2-ethylhexyl) phthalate (DEHP). In: Environmental Science and Pollution Research 24(1), 312-320.

[19] Sugeng, E. J. et al. (2020): Predictors with regard to ingestion, inhalation and dermal absorption of estimated phthalate daily intakes in pregnant women: The Barwon infant study. In: Environment International 139, 105700.

[20] Fink, J. K. (2010): A Concise Introduction to Additives for Thermoplastic Polymers. Hoboken, NJ: Wiley.

[21] Lyche, Jan L. et al. (2015): Human health risk associated with brominated flame-retardants (BFRs). In: Environment international 74, 170-180.

[22] Wang, J. et al. (2018): Legacy and novel brominated flame retardants in indoor dust from Beijing, China: Occurrence, human exposure assessment and evidence for PBDEs replacement. In: Science of the Total Environment 618, 48–59.

[23] Botterell, Z. L. R.; Beaumont, N.; Dorrington, T.; Steinke, M.; Thompson, R.C.; Lindeque, P. K. (2019): Bioavailability and effects of microplastics on marine zooplankton: A review. In: Environmental Pollution 245, 98-110.

[24] Miao, Y.; Wang, R.; Lu, C.; Zhao, J.; Deng, Q. (2016): Lifetime cancer risk assessment for inhalation exposure to di(2-ethylhexyl) phthalate (DEHP). In: Environmental Science and Pollution Research 24(1), 312-320.

[25] EPA (U.S. Environmental Protection Agency). (2020): Integrated Risk Information System https://www.epa.gov/iris

[26] D'Souza, J. M.; Windsor, F. M.; Santillo, D.; Ormerod, S. J. (2020): Food web transfer of plastics to an apex riverine predator. In: Global Change Biology 26(7), 3846-3857.

[27] EPA (U.S. Environmental Protection Agency). (2020): Toxics Release Inventory (TRI) Program (2020): https://www.epa.gov/toxics-release-inventory-tri-program

[28] Franklin Associate. (2011): Cradle-to-gate life cycle inventory of nine plastic resins and four polyurethane precursors. https://plastics.americanchemistry.com/LifeCycle-Inventory-of-9-Plastics-Resins-and-4-Polyurethane-Precursors-Rpt-Only/

[29] Karlsson, T.M.; Arneborg, L.; Brostrom, G.; Almroth, B. C.; Gipperth, L.; Hassellov, M. (2018): The unaccountability case of plastic pellet production. In: Marine Pollution Bulletin 129(1), 52-60.

[30] Lithner, D.; Larsson, A.; Dave, G.; (2011): Environmental and health hazard ranking and assessment of plastic polymers based on chemical composition. In: Science of the Total Environment 409(18), 3309-3324.

AUTHORS DIRECTORY