Implementation of a Self-Enforcing Network to Identify Determinants of the WiFi Quality on German Highspeed Trains

Jonathan Berrisch

Institute for Business and
Economic Studies
University of Duisburg-Essen
Essen, Germany
email: jonathan.berrisch@vwl.uni-due.de

Timo Rammert

Institute for Business and
Economic Studies
University of Duisburg-Essen
Essen, Germany
email: timo.rammert@vwl.uni-due.de

Christina Klüver

Institute for Computer Science and
Business Information
University of Duisburg-Essen
Essen, Germany
email: christina.kluever@uni-due.de

Abstract—**In this paper, we demonstrate how to analyze the WiFi data of the German highspeed trains called InterCityExpress (ICE) on the basis of a neural network. To achieve this, we apply a Self-Enforcing Network with cue validity factors to underline the importance of selected features. It is shown that the quality of the WiFi connection, in terms of the rate of downloads and the latency, can be grouped and explained by just a few determinants. We will show where the network coverage is especially good or bad and suggest ways to improve this quality to enhance the comfort of traveling on the highspeed trains and therefore to possible expand the profits of the operating company.**

Keywords–Self-Enforcing Network (SEN); self-organized learning; cue validity factor; Intelligent data analysis; Industry 4.0 data analysis.

I. Introduction

In recent years, the demand for mobile Internet access grew at a rapid pace. The reasons for this development are diverse. There is ongoing research on how this change affects people in various ways (e.g., [1][2]). Furthermore, there is a growing literature trying to identify the factors which drive the ongoing rise in demand [3][4]. While user habits change, new challenges, and opportunities for many businesses arise.

Technical challenges to manage wireless networks in general [5][6] and in highspeed trains in particular [7][8], are discussed in numerous recent publications, using very different algorithms like k-means, deep- and reinforcement learning, support vector machines, optimization algorithms, Decision Trees, Naive Bayes, to name only a few [5][9][10][11][12][13]. Despite the analysis of technical improvements, the possibility to get other information such as the numbers of train travelers using mobile phone data ([14] for an overview), or how passengers use travel time [15][16] are also of interest.

In an increasingly competitive transportation industry, the existence of free mobile Internet access will be a crucial factor to attract new customers [17]. That is why the biggest German railway company “Deutsche Bahn AG” announced to offer a free WiFi system to all travelers on their highspeed trains in the course of their quality improvement program “Zukunft Bahn” [18]. The main goal of this program is to achieve higher customer satisfaction and hence to expand their profits.

This paper contributes to this sphere by analyzing the supply side of mobile Internet access. To be more specific, we analyze the user experience when using the WiFi network that is provided on German highspeed trains (ICEs) in different regions of Germany.

In particular, the analysis of GPS based data with traditional statistical methods is challenging. While, for example, regression approaches are good at describing continuous linear and nonlinear relationships between variables and are computationally simple, they are not suited for clustering data. Algorithms like k-means clustering are better suited but still require quite strong assumptions like knowledge about the exact number of clusters in the data. Therefore, we utilized a self-organized learning neural network to analyze a dataset containing locations specific WiFi data [19]. The usage of this self-organized learning neural network enabled us to analyze the data with only a minimum of prior assumptions needed and without the need of prior variable selection.

The remainder of this paper is structured as follows. Section II describes the Self-Enforcing Network (SEN). In section III, the technology for providing the WiFi on the trains, the data used and methodology are explained. In section IV, the key results are presented focusing on factors affecting the network coverage; section V concludes.

II. The Self-Enforcing Network (SEN)

SEN is a self-organized learning neural network, developed by the Research Group “Computer-Based Analysis of Social Complexity” (CoBASC). Only the functionalities that are relevant to this analysis are briefly presented. More detailed descriptions of the SEN are found in, e.g., [20][21]. The data, consisting of attributes and objects, are represented in a “semantical matrix” where the rows represent the objects o, and the columns represent the attributes a. The values in the matrix w_{ao} represent the degree of affiliation of an attribute to an object. In this case, the semantical matrix contains the preprocessed real data imported from .csv-files (see below).

The training of the network is done by transforming the (min-max normalized) values of the semantical matrix into the weight matrix of the network according to the learning rule. The most specific for SEN is, that the weight values are not generated at random (as usually in neural networks),

Copyright (c) IARIA, 2019.   ISBN: 978-1-61208-723-8

meaning that the weight matrix displays the real data. In addition, a "cue validity factor" (cvf) [22][20] is introduced to exclude, to dampen, or to increase the importance of selected attributes. The whole learning rule is then defined as follows (see Equation 1) with w being the assigned weight and c being a constant defined as $0 \leq c \leq 1$:

$$\begin{aligned} & w(t+1) = w(t) + \Delta w, \text{and} \\ & \Delta w = c * w_{ao} * cvf_a \end{aligned} \tag{1}$$

As in any neural network, the activation functions play an important role. In SEN several activation functions are at disposal [23]; in this study we have used the logarithmic-linear activation function (LLF), which is defined as follows (see Equation 2) with w_{ij} being the value of the (i,j)th Element of the weight matrix and a_i being the corresponding values of the semantical matrix:

$$a_j = \sum \begin{cases} \log_3(a_i + 1) * w_{ij}, & \text{if } a_i \geq 0. \\ \log_3(|a_i + 1|) * -w_{ij}, & \text{else.} \end{cases} \tag{2}$$

The logarithmic-linear activation function is well suited for our purpose because the dataset includes many extreme observations and also differs much across the different variables. This logarithmic-linear function ensures that those extreme values do not receive weights that would be too high otherwise and thus outweigh the other (smaller) values.

III. Data Description and Analysis

To understand how the WiFi on a German Highspeed Train (ICE) works, the technical background will be discussed here shortly.

Figure 1 gives an overview of the installed hardware on the ICEs to provide the local WiFi. The base of this system is the combined infrastructure of the three big telecommunication companies Telekom, Telefónica, and Vodafone, resulting in a more stable system and better network coverage [18]. The signals (the data) coming from cell towers of all three providers are received by the antenna mounted on top of the train. Then the collected stream of data is processed by the router to merge the different signals and is sent to different access points across the train, from which the travelers can access the mobile Internet by connecting to the WiFi Network. To be recognizable, the Service Set Identifier (SSID) of this Network is always set to "WiFionICE".

A. Data

The dataset called "WiFi on ICE" which is analyzed in this paper is provided by the Deutsche Bahn AG [19]. This dataset consists of about 23.5 million observations with 15 variables. Table I sums up all variables which we have used in our analysis. The sid stores a unique ID of each router used in the trains. Together with the GPS variables, it is possible to match observations to individual train connections. Furthermore, the rate of downloads and the latency are essential factors that influence the user experience. High rates of download are good because the content will be downloaded faster, e.g., a website can be displayed fast, or a movie can be played at higher image quality. High latency values, on the other hand, are undesirable. The latency values describe the time which passes until an initial response from the server is received. Additionally, we included the rate of uploads and data on the number of devices connected to the network to check the consistency of our results and to interpret the latter.

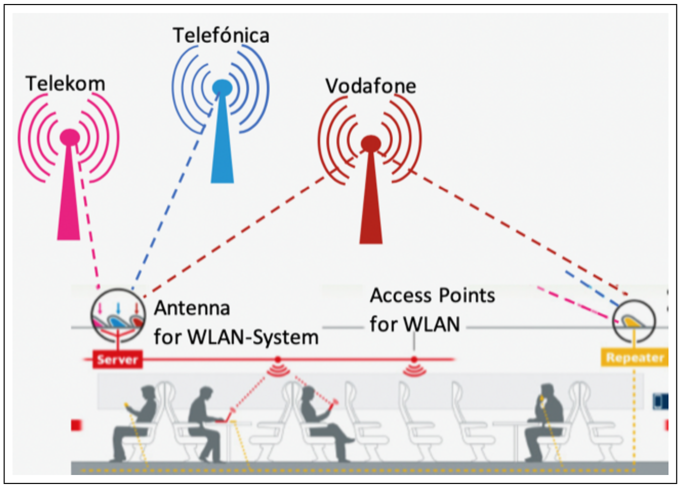


Figure 1. WiFi system on ICE trains [18].

TABLE I. SELECTED VARIABLES.

Variable	*Description*
sid	ID of the X6-router
gps_breite	Latitude
gps_laenge	Longitude
pax_auth	No. of authenticated devices in the local WiFi
pax_total	No. of total devices in the local WiFi
tprx	Rate of downloads (in bytes/s)
tptx	Rate of uploads (in bytes/s)
link_ping	Latency (in ms)
gps_v	Train Speed (in m/s)
gps_richtung	Direction (in degrees)

B. Preprocessing the Data

To get a first impression of the dataset, we utilized R-Studio. R-Studio is an integrated development environment which allows to comfortably program in the statistical programming language R [24].

Because of the huge number of observations looking at them individually was not feasible. Therefore, we looked at first at the number of missing values. In 2 variables, namely *gps_v* and *gps_richtung*, the amount of missing observations exceeds about 10%. In consequence, we omitted all observations that have missing values. This step was necessary to interpret the results later on. Furthermore, we transformed the variable concerning the velocity of the train by multiplying with the factor 3.6 to obtain velocities in kilometers per hour instead of meters per second, which was the original unit of measurement. We excluded observations with an altitude lower than -80 meters and or speed of over 350 kilometers per hour. This is because velocities much larger than this value cannot be reached by an ICE train. Therefore those values have to be seen as results of measurement errors. Lastly, we created a unique identifier which functions as a label. This allows us to easily search for an individual observation of interest when needed. Finally, we are left with 10.1 million observations. Those observations are used to analyze two different train connections across

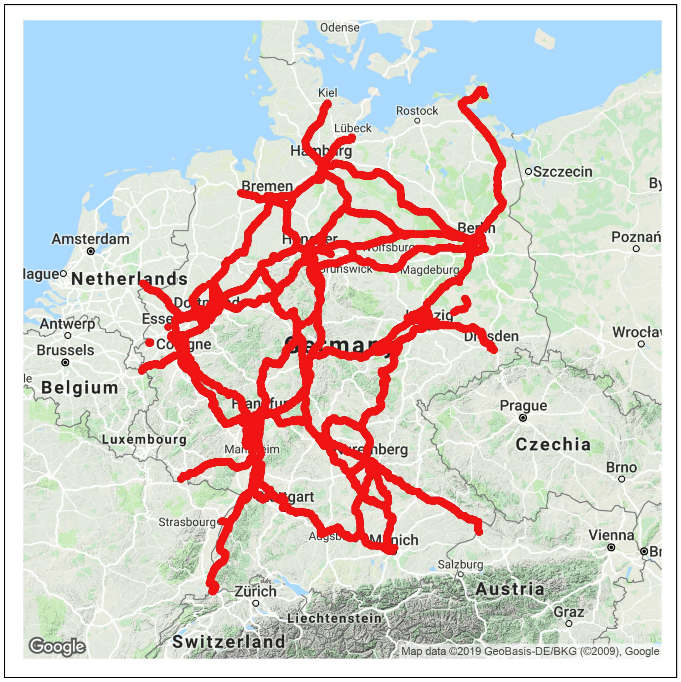


Figure 2. Visualisation of the used dataset.

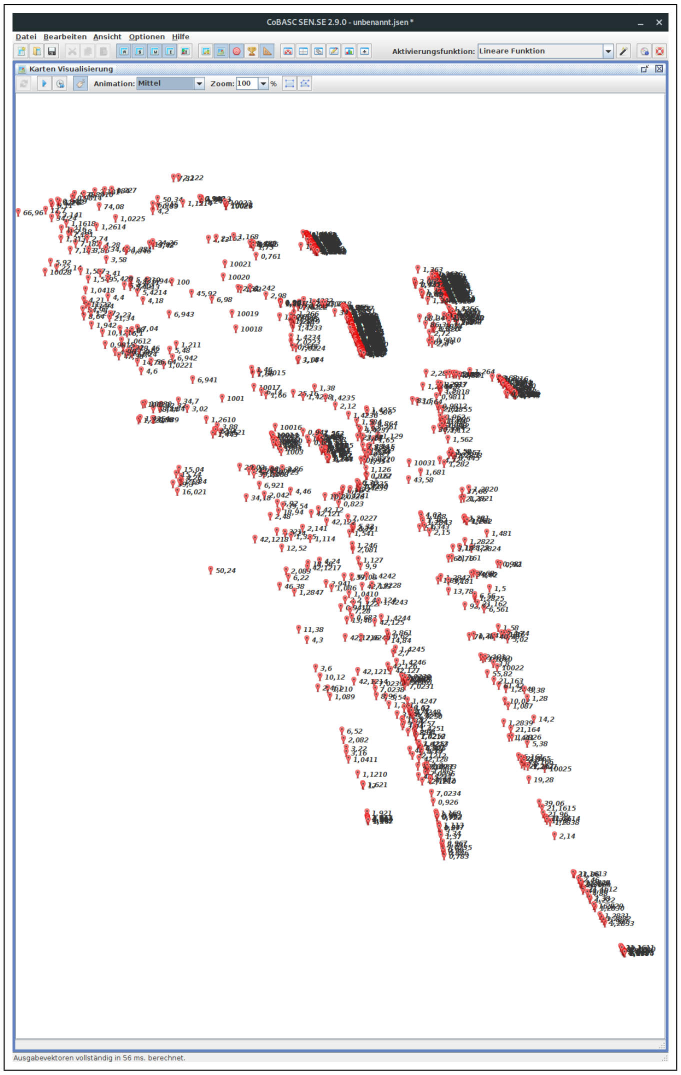
Germany. We chose connections to cover a good amount of the German railway network. The selection of individual train connections was possible by using the *sid* variable. The dataset after preprocessing, which is used in the subsequent analysis, is visualized in Figure 2.

C. Clustering the Data

For clustering the data with regard to the rate of downloads and the latency, the selected data is reduced once more to $1,000$ observations per chosen connection. Those observations are not chosen at random but regularly over the entire observed period of time (e.g., one observation per minute).

Those in total $2,000$ observations in 8 variables are then labeled with their relative latency, and rate of download and afterward imported into Self-Enforcing Network Second Edition (SEN.SE) the tool for applying the neural network and visualizing the results. The first step then was letting the Self-Enforcing Network (SEN) structure the given data without further information or any adjustments. This procedure resulted in the following visualization (see Figure 3).

There are clearly some clusters, and there is some structure in the data. This is good news because it is evidence that the data is not uniformly distributed, i.e., there is information that can be extracted by further analyzing the dataset.



Aktivierungsfunktion: Lineare Funktion

Animation: Mittel

Zoom: 100 %

Ausgabevektoren vollständig in 56 ms. berechnet.

Figure 3. First results of clustering with respect to the normalized latency values.

To steer the clustering towards the variables of interest, the available settings and parameters are set differently. As already mentioned, the Cue Validity Factor (CVF) is a measure for to which extent attributes are associated with a particular category. Therefore one can adjust the importance of an attribute by setting the value of its CVF. If it is set to a high value (i.e., CVF = 1 or possibly even higher), the attribute is highly relevant. If it is set to zero (i.e., CVF = 0), the attribute is not important at all and therefore not considered in the following clustering process.

The value of the CVF for the attributes we want to explain (i.e., the rate of downloads and the latency) is set to the highest value, i.e., 1. The values of the CVF for the other attributes is lowered stepwise until a clear grouping of the objects is reached. A good starting point for setting the CVFs is the correlations between the variables we want to explain and the other covariates. A (in absolute) high correlation indicates that the specific CVF should be set relatively high at the beginning of the process of lowering the values of the CVF for the attributes. The correlations between all the variables in the dataset are visualized in Figure 4. The resulting values of the CVFs are given in Table II. Those adjustments on the SEN then result in the following clustering of the data (see Figure 5). One can see that the data now is ordered in a very different way. Observations with high normalized latency values (red labels) are sorted in the bottom right corner whereas smaller values (i.e., observations with a smaller normalized latency) are sorted in the top left corner (blue labels). The respective

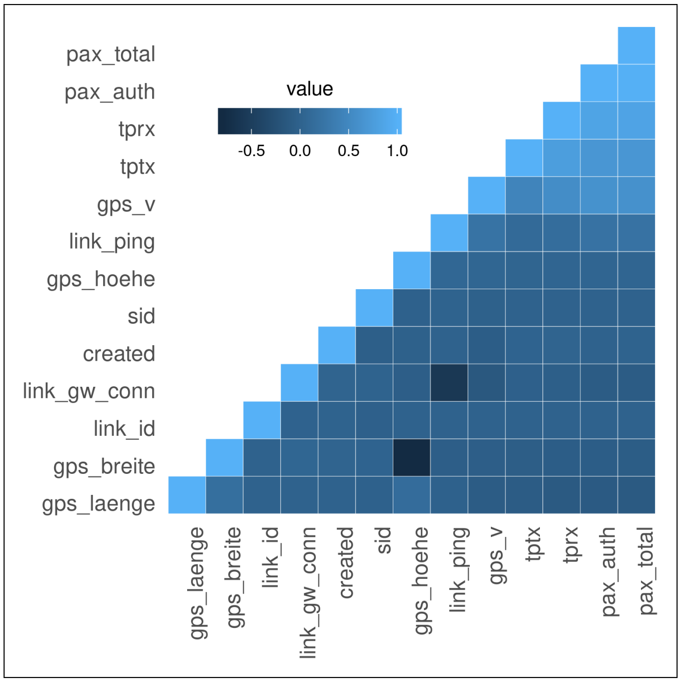


Figure 4. Correlations between the variables.

TABLE II. VALUES OF THE CVFS.

Variable	*Value of CVF to explain latency values*	*Value of CVF to explain rate of downloads*
sid	0	0
gps_breite	0.4	0.3
gps_laenge	0.4	0.3
pax_auth	0.2	0.2
pax_total	0	0
tprx	0	1
tptx	0	0
link_ping	1	0

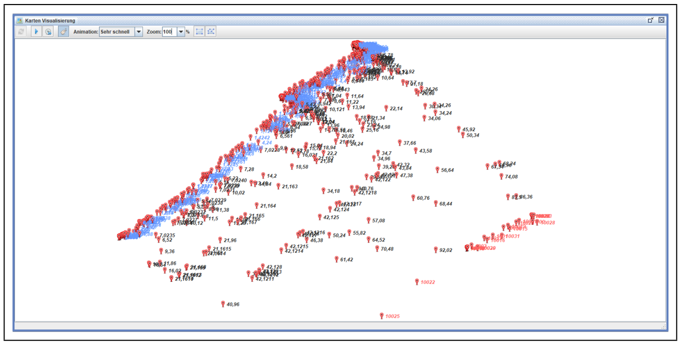
latency values are displayed right next to the corresponding red markers. This clustering, on the one hand, allows us to select regions of observations with very high or low latency or download rates to further analyze them utilizing R. On the other hand one obtains a good impression how the observations are distributed according to their latency values or download rates.

IV. Results

This section presents the key results obtained from the analysis of two different routes across Germany.

A. Berlin - Cologne

ICE line 10, the connection between Berlin in East Germany and Cologne in West Germany, links two important areas of high population density. As the red points in Figure 6 reveal, the rates of downloads are high, especially at the train stations in the major cities. Furthermore, Figure 6 shows that the line segment between Cologne and Hanover is characterized by high rates of downloads whereas the segment between Hanover and Berlin consists of mostly low rates of downloads (indicated by the blue points) with only a few exceptions. Those high rates can be caused by either many active devices in the WiFi or by a bad connection to the Internet.


Karten Visualisierung

Animation: Sehr schnell

Zoom: 100 %

Figure 5. Final clustering with respect to the normalized latency values. Blue labels indicate low values whereas red indicates high latency values.

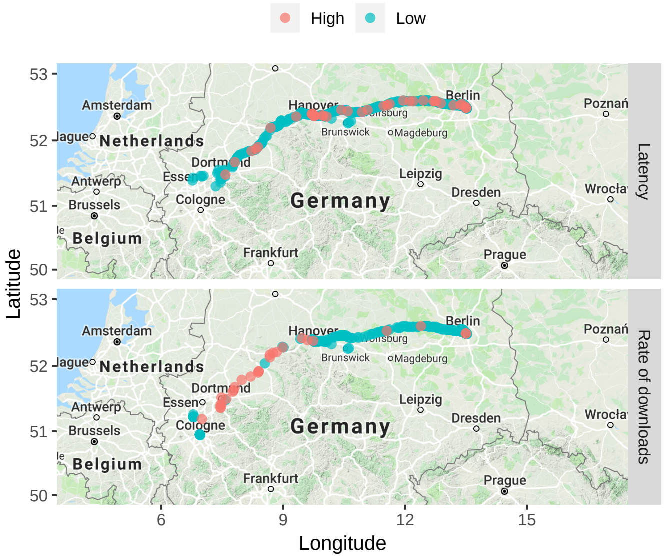


Figure 6. Selected latency values and rates of downloads between Berlin and Cologne.

The latter hypothesis can be evaluated with regard to the latency. Looking at Figure 6 again, it becomes clear, that the latency is far more volatile over the entire length of the connection than the rate of downloads. In particular, between Hanover and Berlin, the latency is constantly relative high explaining the in general lower rates of downloads in this segment. This could possibly be explained by the fact that this region is not highly populated. Therefore the infrastructure of mobile communication may not be that far developed (like for example a larger distance between the cell towers) as for example in the Ruhr area with a much higher population density.

B. Binz - Munich

Figure 7 illustrates the route profile of ICE line 26 connecting Binz in northern Germany and Munich in southern

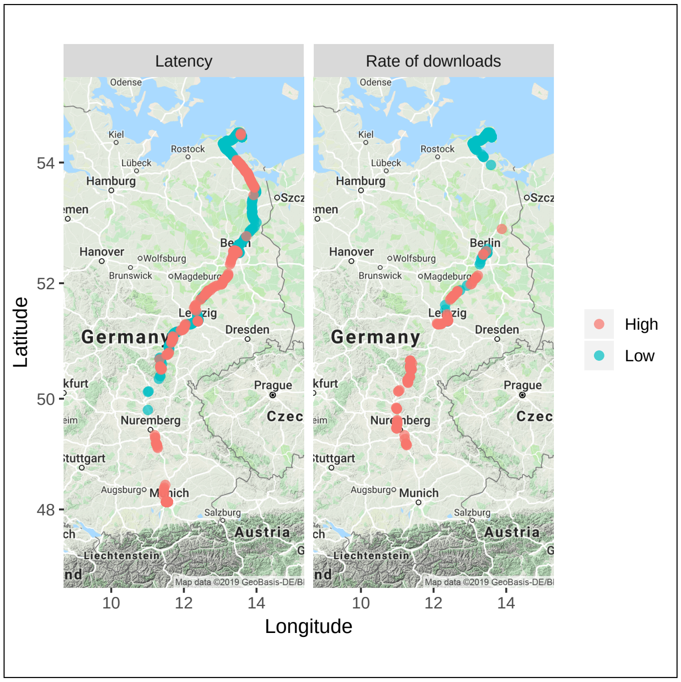


Figure 7. Selected latency values and rates of downloads between Binz and Munich.

Germany. High latency values and high rates of downloads are indicated in red, whereas low ones are blue. It can be seen that on the one hand near Binz, a small coastal town on an island in the Baltic Sea, the rates of downloads are typically low. This comes as no surprise because the population density of this region is quite low. This is also reflected in the usage data, i.e., there are very few devices authenticated to the network. On the other hand, there are quite a few observations with high download rates between Berlin and Nuremberg.

Looking at the latency values, one can notice the low values near Binz. Hence the supply is very good while there is no great demand for it. On the contrary, there is high demand between Berlin and Munich, on average, there are more than 100 devices connected, but the latency is very high most of the time. The latter especially holds between Berlin and Leipzig. Between Leipzig and Nuremberg, the values of the latency are very volatile, which means that the connection to the WiFi might be unstable. Other interesting observations are the high latency values at Munich; they contrast with most observations in other big cities. Unfortunately, the download rates are average at Munich, so it is difficult to assess any further conclusions. At least there is some room for improvement around Munich.

V. Conclusion

In the preceding chapters, we showed how to analyze the WiFi attributes of two selected line-sections of the German railroad network using a SEN. To visualize our results, we used the statistical programming language R. High relevance of our analysis is attributed to the GPS data itself, the download rate, latency and the authenticated devices in the network. While well known statistical approaches suffered fulfilling our requirements, the SEN enabled us to cluster the data without extensive prior model specification.

Our results show a good quality of the WiFi in Northern Germany near Binz. Rather bad quality can be expected when traveling between Berlin and Munich. The identification of areas where the supply of mobile Internet access is good or bad was possible.

We assumed that the user experience using the WiFi mainly depends on the latency and the download rate. To prove this, one could analyze the correlation between data on the user experience and our results. Unfortunately, we were not able to gather data on the user experience, so this remains an exciting area of subsequent research. Furthermore, subsequent research could utilize our methodology to analyze which factors determine the connection between trains and cell towers by including additional data like micro deployment data of the antennas (e.g., the distance between antennas, signal strength, load, etc.). Unfortunately, micro-deployment data is rarely publicly available. This would allow exploiting further relationships between the quality of the WiFi on the ICEs and the mobile network in Germany. Furthermore, one could explore the relationships between train delays and network usage.

While our results cannot be generalized because they are specific for Germany, the generalization of our methodology should be possible. Our analysis mainly relies on the assumption that WiFi access is positively correlated to the customer satisfaction of railway companies. If this assumption holds, and proper data is available, one can conduct the same analysis for other countries.

References

[1] P. Holicza and E. Kadëna, “Smart and Secure? Millennials on Mobile Devices,” Interdisciplinary Description of Complex Systems, vol. 16, no. 3-A, 2018, pp. 376–383, URL: https://ideas.repec.org/a/zna/indecs/v16y2018i3-ap376-383.html [retrieved: 2019-05-14].

[2] J. Martins, C. Costa, T. Oliveira, R. Gonçalves, and F. Branco, “How smartphone advertising influences consumers’ purchase intention,” Journal of Business Research, vol. 94, jan 2019, pp. 378–387, URL: https://linkinghub.elsevier.com/retrieve/pii/S0148296317305507 [retrieved: 2019-05-14].

[3] M. C. Enache, “E-commerce Trends,” ”Dunarea de Jos”, no. 2, 2018, pp. 67–71.

[4] S. Abel, L. Mutandwa, and P. L. Roux, “A Review of Determinants of Financial Inclusion,” International Journal of Economics and Financial Issues, vol. 8, no. 3, 2018, pp. 1–8, URL: https://ideas.repec.org/a/eco/journ1/2018-03-1.html [retrieved: 2019-05-14].

[5] Y. Fu, S. Wang, C.-X. Wang, X. Hong, and S. McLaughlin, “Artificial Intelligence to Manage Network Traffic of 5G Wireless Networks,” IEEE Network, vol. 32, no. 6, nov 2018, pp. 58–64, URL: https://ieeexplore.ieee.org/document/8553655/ [retrieved: 2019-05-14].

[6] H. Zhang and L. Dai, “Mobility Prediction: A Survey on State-of-the-Art Schemes and Future Applications,” IEEE Access, vol. 7, 2019, pp. 802–822, URL: https://ieeexplore.ieee.org/document/8570749/ [retrieved: 2019-05-14]. [Online]. Available: https://ieeexplore.ieee.org/document/8570749/

[7] L. Li et al., “A measurement study on multi-path TCP with multiple cellular carriers on high speed rails,” in Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication - SIGCOMM ’18, 2018, pp. 161–175.

[8] M. Munjal and N. P. Singh, “Group mobility by cooperative communication for high speed railway,” Wireless Networks, jan 2019, pp. 1–10, URL: http://link.springer.com/10.1007/s11276-018-01923-2 [retrieved: 2019-05-14].

[9] R. Cheng, Y. Song, D. Chen, and X. Ma, “Intelligent Positioning Approach for High Speed Trains Based on Ant Colony Optimization and Machine Learning Algorithms,” IEEE Transactions on Intelligent

Transportation Systems, 2018, pp. 1–10, URL: https://ieeexplore.ieee.org/document/8527682/ [retrieved: 2019-05-14].

[10] Y. Bi, J. Zhang, Q. Zhu, W. Zhang, L. Tian, and P. Zhang, "A Novel Non-Stationary High-Speed Train (HST) Channel Modeling and Simulation Method," IEEE Transactions on Vehicular Technology, vol. 68, no. 1, jan 2019, pp. 82–92, URL: https://ieeexplore.ieee.org/document/8542786/ [retrieved: 2019-05-14].

[11] C. Wu, Y. Wang, and Z. Yin, "Realizing Railway Cognitive Radio: A Reinforcement Base-Station Multi-Agent Model," pp. 1–16, 2018, URL: https://ieeexplore.ieee.org/document/8424487/ [retrieved: 2019-05-14].

[12] Q. Mao, F. Hu, and Q. Hao, "Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey," pp. 2595–2621, 2018, URL: https://ieeexplore.ieee.org/document/8382166/ [retrieved: 2019-05-14].

[13] K. Sabanci, E. Yigit, D. Ustun, A. Toktas, and M. F. Aslan, "WiFi Based Indoor Localization: Application and Comparison of Machine Learning Algorithms," in 2018 XXIIIrd International Seminar/Workshop on Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory (DIPED). IEEE, sep 2018, pp. 246–251, URL: https://ieeexplore.ieee.org/document/8543125/ [retrieved: 2019-05-14].

[14] A. Ø. Sørensen, N. O. E. Olsson, M. M. Akhtar, and H. Bull-Berg, "Approaches, technologies and importance of analysis of the number of train travellers," Urban, Planning and Transport Research, vol. 7, no. 1, jan 2019, pp. 1–18, URL: https://www.tandfonline.com/doi/full/10.1080/21650020.2019.1566022 [retrieved: 2019-05-14].

[15] J. Tang, F. Zhen, J. Cao, and P. L. Mokhtarian, "How do passengers use travel time? A case study of Shanghai Nanjing high speed rail," Transportation, vol. 45, no. 2, mar 2018, pp. 451–477, URL: http://link.springer.com/10.1007/s11116-017-9824-9 [retrieved: 2019-05-14].

[16] B. Wang and B. P. Loo, "Travel time use and its impact on high-speed-railway passengers' travel satisfaction in the e-society," pp. 1–13, may 2018, URL: https://www.tandfonline.com/doi/full/10.1080/15568318.2018.1459968 [retrieved: 2019-05-14].

[17] U. Kluge, A. Paul, M. Urban, and H. Ureta, "Assessment of Passenger Requirements Along the Door-to-Door Travel Chain." Springer, Cham, 2019, pp. 255–276, URL: http://link.springer.com/10.1007/978-3-319-99756-8{_}17 [retrieved: 2019-05-14].

[18] Deutsche Bahn AG, "Zukunft Bahn: Statusbericht Nordrhein-Westfalen 2016," DB AG, Tech. Rep., 2018, URL: https://www.deutschebahn.com/resource/blob/1296202/c1423d8966c791c726550fe9e98ecf20/20161129-Praeesentation-Zukunft-Bahn-NRW-data.pdf [retrieved: 2019-05-14].

[19] Deutsche Bahn, "Dataset: wifi-on-ice," 2019, URL: https://data.deutschebahn.com/dataset/wifi-on-ice [retrieved: 2019-05-14].

[20] C. Klüver, "Steering clustering of medical data in a Self-Enforcing Network (SEN) with a cue validity factor," in 2016 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, dec 2016, pp. 1–8, URL: http://ieeexplore.ieee.org/document/7849883/ [retrieved: 2019-05-14].

[21] C. Klüver, J. Klüver, and D. Zinkhan, "A self-enforcing neural network as decision support system for air traffic control based on probabilistic weather forecasts," in Proceedings of the International Joint Conference on Neural Networks, vol. 2017-May. IEEE, may 2017, pp. 729–736, URL: http://ieeexplore.ieee.org/document/7965924/ [retrieved: 2019-05-14]. [Online]. Available: http://ieeexplore.ieee.org/document/7965924/

[22] E. Rosch and C. B. Mervis, "Family resemblances: Studies in the internal structure of categories," Cognitive Psychology, vol. 7, no. 4, oct 1975, pp. 573–605, URL: https://www.sciencedirect.com/science/article/pii/0010028575900249 [retrieved: 2019-05-14].

[23] C. Klüver, "Self-Enforcing Networks (SEN) for the development of (medical) diagnosis systems," in Proceedings of the International Joint Conference on Neural Networks, vol. 2016-Octob. IEEE, jul 2016, pp. 503–510, URL: http://ieeexplore.ieee.org/document/7727241/ [retrieved: 2019-05-14].

[24] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019, URL: https://www.R-project.org/ [retrieved: 2019-05-14].