

Monitoring Data Stream Reliability in Smart City Environments

Daniel Kuemper, Thorben Iggena, Marten Fischer and Ralf Toenjes

*Lab for RF-Technology and Mobile Communications
University of Applied Sciences Osnabrueck, Osnabrueck, Germany
Email: d.kuemper@hs-osnabrueck.de*

Abstract—Reliable information processing is an indispensable task in Smart City environments. Heterogeneous sensor infrastructures of individual information providers and data portal vendors tend to offer a hardly revisable information quality. This paper proposes a correlation model-based monitoring approach to evaluate the plausibility of smart city data sources. The model is based on spatial, temporal, and domain dependent correlations between individual data sources. A set of freely available datasets is used to evaluate the monitoring component and show the challenges of different spatial and temporal resolutions.

Keywords—smart city, monitoring, plausibility, traffic data, time series, spatio-temporal, reasoning

I. INTRODUCTION

The quality of experience using smart city applications highly depends on the availability of appropriate, accurate, and trustworthy data. This includes the availability of necessary data sources as well as proving their plausibility. The reliability of the extracted sensor information of external data sources has to be monitored permanently during runtime to obtain satisfying application functionality and user experience. Smart city infrastructures are using a variety of different information sources facing a divergent trustworthiness of information providers and their sensor equipment. In contrast to common simplified IoT sensor infrastructures, cities often utilise aggregated and reasoned information. Traffic data, for example, often rests upon only a subset of cars that are measured and is interpolated to offer a view of the whole situation. Due to a frequent unavailability of precise sensor data and a missing ground truth, there is a high need for evaluating data source reliability and determining the trustworthiness. Therefore, this work proposes a correlation model - based monitoring approach to evaluate smart city data sources.

The remainder is structured as follows: Section 2 presents the state of the art whereas section 3 describes a model for the definition of data stream correlations and available datasets. Section 4 describes the evaluation concept of the monitoring approach whereas a performance evaluation is discussed in section 5. The paper concludes in section 6.

II. STATE OF THE ART

The assessment of Quality of Data can basically be evaluated in 5 common dimensions: Completeness, Correctness, Concordance, Plausibility and Currency. [1] provides a table of different terms used to describe one of the dimensions of data quality. Furthermore, they provide a mapping between data quality dimensions and data quality assessment methods. [2] introduced Sieve, a framework to flexibly express quality assessment methods and fusion methods. This module leveraged user-selected meta-data as quality indicators to produce quality assessment scores through user-configured scoring functions. In their work they use the "fit for use" definition, which is a subjective way of looking at data quality, since it depends on the data consumer. In this paper, we describe a method of assessing the plausibility of data in an objective way for better comparability. The STAR-CITY project [3] describes a system for semantic traffic analytics. Based on various heterogeneous data sources (e.g., Dublin bus activity, events in Dublin city) their system is able to predict future traffic conditions with the goal to make traffic management easier and to support urban planning.

One of the major challenges in the assessment of a quality metric to sensory smart city data is the lack of ground truth. This problem is well-known in the image processing domain when trying to rate the quality of a single picture without any reference. The task is often described as "blind image quality assessment" [4] or No-Reference Image Quality Assessment (NR-IQA) [5]. The concept refers to an automatic quality assessment of an image using an algorithm such that the only information that the algorithm receives is the image whose quality is being assessed. To get an objective quality metric NR-IQA analyses the sharpness of edges or the noise levels. While these approaches can be used to determine the quality of data they are unfit to make a statement about the plausibility of the information, both in image processing and sensor networks. The authors of [6] and [7] developed and evaluated a concept for the assessment of node trustworthiness in a network (vehicular ad-hoc networks VANETs to be precise) based on data plausibility checks. For this, they employed a Bayesian filter consisting of a predict/update cycle. They propose that every node performs a plausibility

check to identify malicious nodes sending faulty data. Like in this work they use similar data sources in order to find “witnesses” for a given sensor reading. The authors in [8] propose three different approaches to deal with a missing ground truth in social media: spatiotemporal, causality, and outcome evaluation. Their concept to use spatiotemporal evaluation to predict future behaviour of humans is similar to our approach disregarding that we evaluate past events. Prior work of the authors emphasised the importance of an appropriate distance model reflecting infrastructure, e.g., roads, and physics, i.e. traffic or air movements[9]. The proposed approach in this work refines the state of the art by utilising sensor and domain independent correlation models whilst incorporating knowledge of the city infrastructure to evaluate data stream plausibility.

III. STREAM CORRELATION MODEL

The analysis of isolated data streams is not a feasible solution in determining the correctness of outliers and suspicious stream behaviour. This paper proposes a correlation model between data streams to determine data plausibility with the dependency of relevant data streams. The model describes expected relations between individual information sources and their mutual impact. Thereby, the model does not necessarily describe inevitable effects between data streams. Therefore, when a traffic sensor *A* reports an average car count of 0 vehicles per minute, a data source *B* that reports traffic jams out of different sensor data can be used to verify that the sensor *A* is working correctly.

A. Datasets

For a basic evaluation of the proposed approach the datasets, described in Table I, have been used. Open Data Aarhus (ODAA)¹ is an open data portal for the Aarhus municipality and allows access to various datasets, e.g., traffic and cultural event data. The traffic data consists of 449 deployed traffic sensors in the city, which report information about the vehicle count and average speed on the defined sensor location. Only a subset of driving cars is measured since these traffic sensors only measure activated WiFi/Bluetooth communication devices. During the City-Pulse Project [10] this data has been collected for more than two years and is available as a semantically annotated dataset².

The second used data source TomTom³ allows access to a set of developer APIs that deliver traffic information. The Traffic Flow API provides information about the current average vehicle speed and the typical vehicle speed of a requested area. The Traffic Incidents API delivers information about registered events, categorised in 4 Levels: 1 Slow Traffic, 2 Queuing Traffic, 3

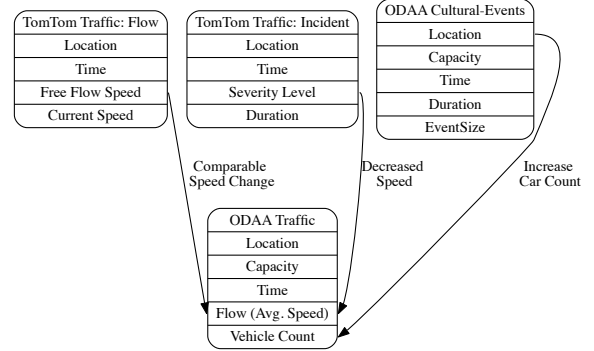


Figure 1. Data Stream Relations

Stationary Traffic, 4 Closed Road. This traffic information is generated from movement of mobile devices (navigation apps, in-dash navigation, phones [11] [12]).

Both data sources are measuring just a subset of the cars that are driving on the road of interest. Therefore, they are not able to deliver distinct and precise information of all vehicles in the city. This paper shows that a cross-evaluation of both datasets, which are generated out of two independent sensor sets, is a useful measure to determine their reliability.

The estimated correlation model between the described data streams is illustrated in Figure 1. It visualises the expected impact of ODAA Cultural-Events and TomTom Traffic Flow/Incident events on the ODAA Traffic data. The following section describes the applied methodology for the evaluation.

IV. COMPOSITE MONITORING METHODOLOGY

The Composite Monitoring [13] evaluates the plausibility of individual events, thus evaluating the data streams causing the event. This is realised in the domain of large sensor networks, e.g., those deployed in a smart city, where no ground truth is available. The monitoring employs model-based analysis of different spatially and temporally related sensor values. The model-based approach allows detecting outliers in sensor readings that are caused by defect sensors and cannot be explained by similar information of related sensors. For example, a traffic jam can be detected by traffic sensors reporting extensively slower traffic speeds. This can be validated by an analysis of consecutive traffic sensors on a road. Therefore, the Composite Monitoring inspects historic time series of various dependent sensor streams.

To evaluate the plausibility of a reported event, in the first step, nearby data streams are identified. Based on the category (e.g., Temperature, Parking, Traffic) of these sources further sensors, which are located nearby are selected. According to the used model, the event should also affect neighbouring sensors and cause a similar behaviour. Real world events exhibit a typical spatial propagation that can be modelled. For example, traffic propagates along the roads whereas noise propagates in every direction. Figure 2 shows the Voronoi cells for ODAA traffic flow sensors

¹<https://www.odaa.dk>

²<http://www.ict-citypulse.eu/page/content/tools-and-data-sets>

³<http://developer.tomtom.com>

Table I
EXAMINED DATASETS

Source	Dataset	Spatial Resolution	Provision	Frequency	Information	Sensor	Reasoning
ODAA	Traffic	Road Segment	Periodic	1/5min	Flow Vehicle Count	Speed (km/h) Cars/5min	Agg.(Mean) Agg.(Sum)
	Cultural-Events	Event Location	Periodic	1set/day	Capacity	Max. Visitors	Agg.(Sum)
TomTom	Traffic Flow	Road Segment	Periodic	1/30min	Current Speed Free Flow Speed	km/h	Agg.(Mean)
	Traffic Incident	Coordinate/Direction	Event-Based	-	Set of Incidents	Severity (4 levels)	Reasoned

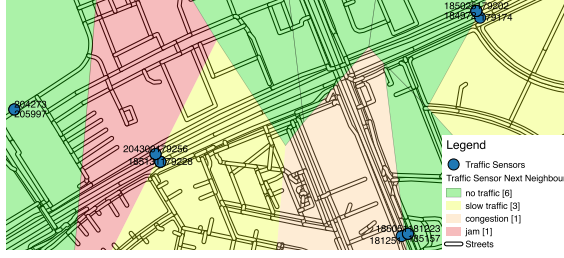


Figure 2. Voronoi Diagram - Depicting the Nearest Traffic Flow for Every Street Segment

and illustrates that the nearest traffic sensor is unlikely to represent the current flow of the illustrated street segments.

Figure 3 describes the evaluation process of the Composite Monitoring. The goal is to determine a correctness value C_e for the event e . A set of correlating data streams S_e is used as validation source. A set of sensor specific validation functions V_s is used to compute the plausibility of the event e . The figure shows the four phases of the evaluation process:

- 1) Determine relevant sensors in the set of all streams S . Find spatially correlated streams by using a suitable distance model M_d , which describes the means of propagation of the event (e.g. air or street)
- 2) Determine the temporal distance by analysing the direction d of expansion, propagation velocity v , and range r of the impact as function of M_d
- 3) Compute the correlations between streams and the event e by applying
 - a) V_s as the set of validation functions for event e and each stream $s \in S_e$
 - b) τ_s as the set of temporal direction (defines if the change in s is a result of e or the cause for e)
- 4) Analyse partial correlation values to evaluate the correctness by using a set of weights W_s for each stream $s \in S_e$ and a combination function \sum , e.g. min, mean.

As a result, we get the tuple

$$C_e = (S, M_d, d, v, r, V_s, \tau_s, W_s, \sum)$$

that defines the evaluation Model.

A. Decomposition of Stream Data

Since continuous smart city data is often based on seasonal patterns and appears as noisy time-based datasets the

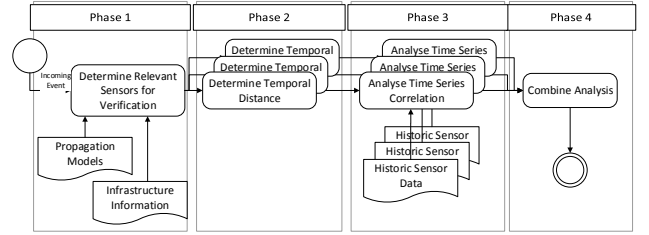


Figure 3. Composite Monitoring Process

ability to separate trend, seasonality and irregular components allows an interpretation of the current situation. For example, when there are no cars measured on a small street at night, it does not typically mean that there is a traffic jam. Therefore the Composite Monitoring applies a decomposition of the time series into additive seasonal, and irregular components using moving averages. The trend of the time series is a useful indicator for the needed length of the analysed data stream period but is not utilised in the for an incident evaluation.

S_t The *Seasonal Component* at time t reflects seasonally repeated variation. It exists when a time series is influenced by seasonal factors for a fixed and known period (e.g., daily traffic pattern like morning and evening rush hours).

I_t The *Irregular Component* at time t , that describes random, irregular influences, which counter regular patterns and therefore have the most significance for current regular and irregular sensor situations. It represents the residuals or remainder of the time series after the other components have been removed.

Hence, the time series using an additive model can be thought of as: $y_t = S_t + I_t$. Figure 4 shows the exemplary decomposition of an ODAA traffic sensor time series (Raw) with an illustrated traffic event.

B. Smoothing Noisy Data

Since the available ODAA data, which only measures a subset of bypassing cars every 5 minutes, shows a high noise between different consequent measurements a smoothing filter function was applied to get data that is more comparable. The discrete cosine transform (DCT) expresses a finite sequence of data points as a sum of cosine waves

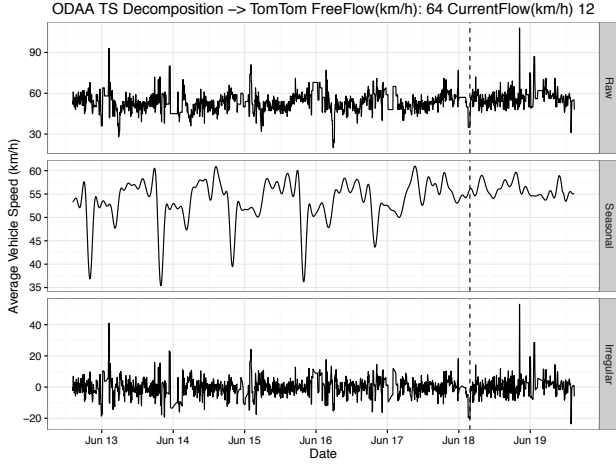


Figure 4. Decomposition of Timeseries, Event at Vertical Dashed Line

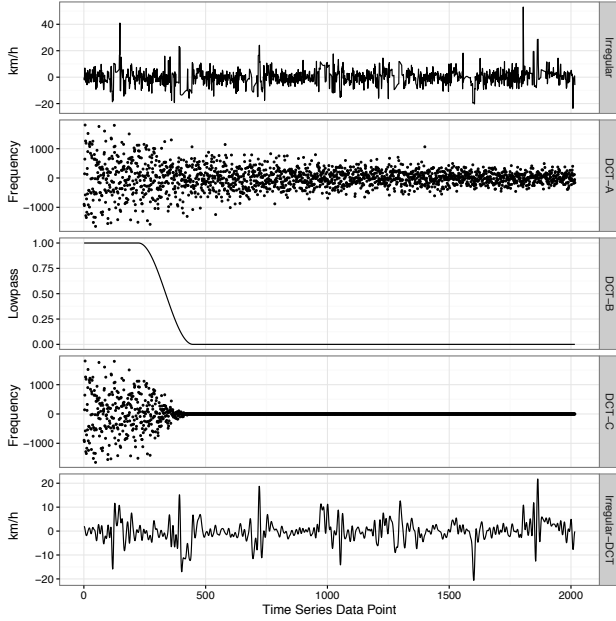


Figure 5. DCT Filter to Remove High Noise on the Irregular Component of the Decomposed Time Series

with different frequencies and amplitudes. The implemented discrete cosine transform filter depicted in Figure 5 uses the DCT to convert a signal (Irregular) to an ordered sequence of frequencies and associated amplitudes (DCT-A). In the frequency domain, the sequence is multiplied with a low pass (DCT-B) to remove high frequencies (DCT-C). The inverse DCT transforms the low pass filtered sequence back to the time domain resulting in a smoothened signal (Irregular-DCT).

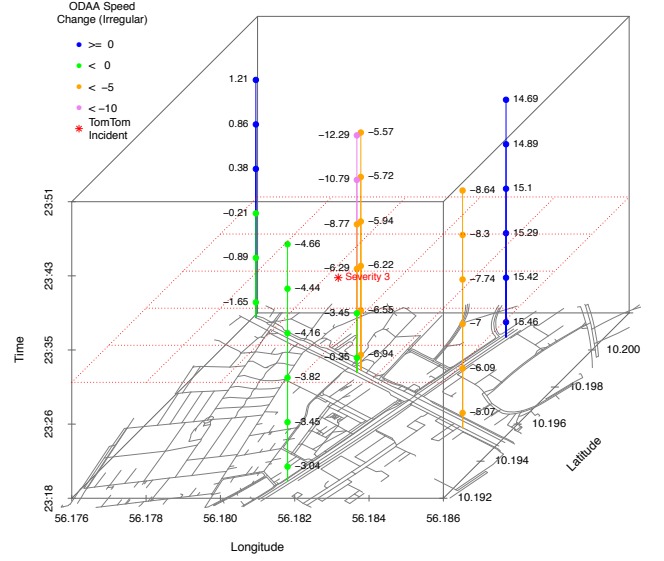


Figure 6. Example of Space-Time Resolution of Sensors and Events

C. Space Time Resolution Example

The following Figure 6 illustrates the appearance of a TomTom Incident report in the city of Aarhus combined with nearby ODAA Traffic data streams. The x-axis and z-axis show the longitude and latitude of the information source. The y-axis shows the time of measurement. At the origin of the z-axis, a grey plot of the street segments [14] is visualised. The TomTom Traffic Incident is plotted as a red star. To highlight the time of the event a grid is projected over the whole area. The severity level of the incident is 3 *Stationary Traffic*, which is defined as a strong reduction of speed. To evaluate the event the ODAA sensors inside a 200m Radius have been selected. Six weeks of sensor readings prior to the event have been used to decompose the time series data. After the decomposition, for each of the investigated ODAA sensors the irregular component of six consecutive sensor readings have been analysed, 3 readings before and 3 readings after the event. Since the sensor provides new readings every 5 minutes this covers 30 minutes around the event. These values are plotted on the map at the measurement location of the sensors and visualise irregular component of the average speed.

The analysis shows a significant speed drop during the TomTom Event on several sensors that measure the same direction inside the street-crossing area.

D. Event-Based Results

The experiment in the previous section shows the evaluation of a single incident. This section evaluates the plausibility of 8607 incidents, that have been captured by the authors with the TomTom API for Aarhus, against the ODAA sensor data set (six weeks time series data for up to

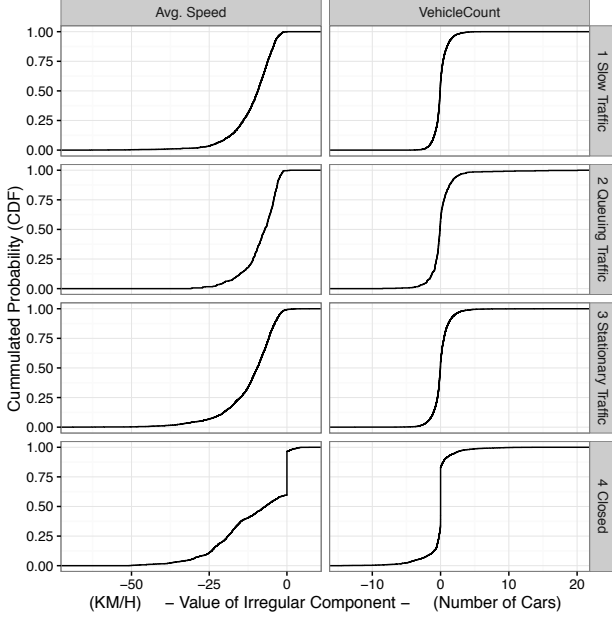


Figure 7. ODAA Datasets Avg. Speed (Irregular Component) and Vehicle Count (Raw Data) Change During TomTom Traffic Incidents

10 sensors within a maximum distance of 200 meters). The overall results show that for more than 75% of the severity level 4-closed the raw data of a nearby ODAA sensor measured a vehicle count of 0 cars during the event period. Figure 7 shows the cumulative density function (CDF), of the irregular component of the ODAA during the TomTom incidents. It shows that for reported events most vehicle speed readings are slower than described by the seasonal component S_t , which means that the irregular component $I_t < 0$. The change of the vehicle count does not clearly reflect the event. The vehicle count can drop because of a traffic jam or a traffic jam can be the result of the road being overloaded.

E. Continuous Measurement Result

After the discussion of event-based sensor readings in the last section, this section deals with measurements on a continuous scale. Therefore TomTom Traffic Flow data has been evaluated against ODAA traffic flow measurements. For every TomTom measurement that showed a severe slow-down of more than 15km/h compared to the free flow speed an evaluation has been triggered: Six weeks time series data for up to 10 sensors within a maximum distance of 200 meters have been analysed and their minimum values of the irregular component have compared to the TomTom measurement. Figure 8 shows the heatmap distribution of 28306 pairwise comparisons of the two data sources. The colour in the heatmap shows the density of measurement points. The figure depicts that there is no clear linear correlation between the two data sources. 95.5% of ODAA comparisons also have an irregular component value $I_t \leq 0\text{km/h}$, which is

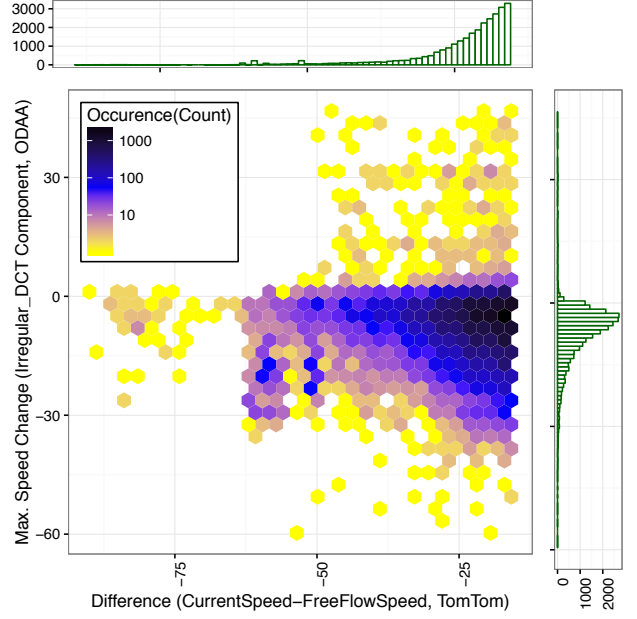


Figure 8. Heatmap of Measured Value Pairs of Random Part Change (ODAA) Against Traffic Flow Reports (15km/h Speed Decrease)

indicating a decrease in speed. These measurements confirm the plausibility of the TomTom Flow data, although there is no direct correlation between measured values.

V. PERFORMANCE EVALUATION

Evaluation of sensor data is a time critical task since its usefulness and validity is not unlimited. Furthermore, the number of sensor sources is growing rapidly with the result that future analysis has to be scalable and support a huge number of sensors in parallel. Figure 9 shows the performance measurement of the previously presented time series decomposition and comparison/correlation calculation to a given value or event. It shows boxplots for 10 repetitions of the experiments. The boxplot whiskers show the lowest duration still within 1.5 Inter-Quartile Range (IQR) of the lower quartile and the highest duration still within 1.5 IQR of the upper quartile. Outliers are drawn as a dot. The x-axis shows the numbers of sensors that were used for the evaluation. The y-axis shows the duration of the evaluation. The colour of the box illustrates 3 different time-spans that were used for the time series decomposition that has been used for the evaluation. The figure shows a basically linear time increase for a rising number of sensors. However, the graphs are not consistently linear since the algorithm has been parallelized in 11 threads for multiple sensors⁴. Outliers mainly depend on active system tasks that have been running in parallel. The performance evaluation shows that e.g., an

⁴Evaluation System: Intel Core i7-5820K, 3.6ghz, 32GB RAM, SSD, PostgreSQL database as historic data source

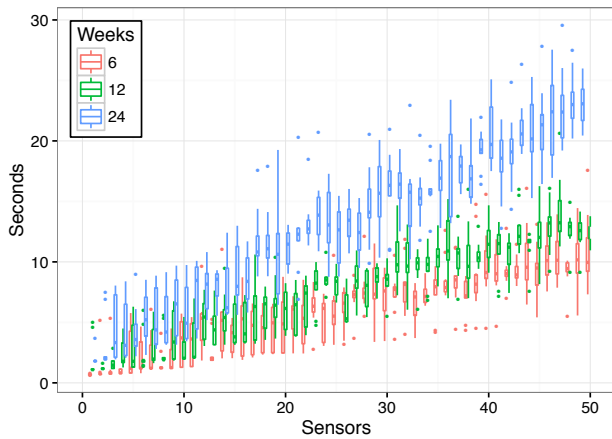


Figure 9. Exemplarily Processing Times for Loading, Decomposing and Comparing Time Series to Events

evaluation of 10 correlating sensors, regarding 12 weeks of historic data and decomposing it for a correlation analysis can be approached in usually less than 5sec. This leads to a feasible utilisation in a live environment for the evaluation of suspicious data. Furthermore a caching of time series (e.g., once a week) allows the hardware setup to substitute processing power with memory.

VI. CONCLUSION

This paper discusses measures for ensuring and increasing the reliability of smart city applications by employing a monitoring approach for correlating data streams. During run-time, the monitoring evaluates the plausibility of sensor observations. Composite Monitoring, can be triggered by detected incidents or suspicious sensor readings in data streams. It provides a correlation dependency model-based approach integrating new evaluation schemes. This approach supports appropriate spatiotemporal distance measures to achieve an efficient monitoring of relevant correlating data. In conclusion, the suggested framework provides methods to cope plausibility analysis for heterogeneous data sources in smart city applications, and in addition considers that the sensors become unreliable over time by approaching a continuous time series evaluation. As a countermeasure, several actions to identify and react on varying information qualities have been investigated and integrated into the CityPulse framework. In the future, we plan to investigate an open platform for a drag and drop evaluation approach, which allows to apply the concept to different domains and allows easy utilisation of city infrastructure knowledge.

ACKNOWLEDGMENT

This work is supported by the European Union 7th FP Project CityPulse under grant agreement n. 609035

REFERENCES

- [1] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.
- [2] P. N. Mendes, H. Mühleisen, and C. Bizer, "Sieve: Linked data quality assessment and fusion," in *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, ser. EDBT-ICDT '12. New York, NY, USA: ACM, 2012, pp. 116–123.
- [3] F. Lécué, S. Tallevi-Diotallevi, J. Hayes, R. Tucker, V. Bicer, M. Sbodio, and P. Tommasi, "Smart traffic analytics in the semantic web with star-city: Scenarios, system and lessons learned in dublin city," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 27, pp. 26–33, 2014.
- [4] X. Li, "Blind image quality assessment," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, 2002, pp. I-449–I-452 vol.1.
- [5] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec 2012.
- [6] N. Bissmeyer, S. Mauthofer, K. M. Bayarou, and F. Kargl, "Assessment of node trustworthiness in vanets using data plausibility checks with particle filters," in *Vehicular Networking Conference (VNC), 2012 IEEE*, Nov 2012, pp. 78–85.
- [7] N. Bissmeyer, J. Njeukam, J. Petit, and K. M. Bayarou, "Central misbehavior evaluation for vanets based on mobility data plausibility," in *Proceedings of the Ninth ACM International Workshop on Vehicular Inter-networking, Systems, and Applications*, ser. VANET '12. New York, NY, USA: ACM, 2012, pp. 73–82.
- [8] R. Zafarani and H. Liu, "Evaluation without ground truth in social media research," *Communications of the ACM*, vol. 58, no. 6, pp. 54–60, 2015.
- [9] R. Toenjes, D. Kuemper, and M. Fischer, "Knowledge-based spatial reasoning for iot-enabled smart city applications," in *2015 IEEE International Conference on Data Science and Data Intensive Systems*. IEEE, 2015, pp. 736–737.
- [10] D. Puiu, P. Barnaghi *et al.*, "Citypulse: Large scale data analytics framework for smart cities," *IEEE Access*, vol. 4, pp. 1086–1108, 2016.
- [11] M. Clements, "Detecting dangerous locations in the road network with jam tail warnings," Oct. 2015, 00151.
- [12] R.-P. Schäfer, "Iq routes and hd traffic: technology insights about tomtom's time-dynamic navigation concept," in *Proceedings of the ACM SIGSOFT symposium on The foundations of software engineering*. ACM, 2009, pp. 171–172.
- [13] D. Kuemper, T. Iggena, M. Bermudez-Edo, D. Puiu, M. Fischer, and F. Gao, "Measures and metrics for reliable information processing, citypulse deliverable 4.1," Tech. Rep., 2015.
- [14] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008.