Exploring the Quality of Dynamic Open Government Data for Developing Data Intelligence Applications: The Case of Attica Traffic Data

ARETI KARAMANOU*, Centre for Research & Technology Hellas (CERTH), Greece and University of Macedonia, Department of Business Administration, Greece

PETROS BRIMOS*, University of Macedonia, Department of Business Administration, Greece

EVANGELOS KALAMPOKIS*, Centre for Research & Technology Hellas (CERTH), Greece and University of Macedonia, Department of Business Administration, Greece

KONSTANTINOS TARABANIS*, Centre for Research & Technology Hellas (CERTH), Greece and University of Macedonia, Department of Business Administration, Greece

Dynamic data (including environmental, traffic, and sensor generated data) were, recently, recognised as an important part of the Open Government Data (OGD) movement. These data are of vital importance in the development of data intelligence applications. For example, various business applications exploit traffic data to predict, e.g., traffic demand and an estimated time of arrival. However, this type of data is inherently vulnerable to data quality errors produced by, e.g., failures of sensors and network faults. The objective of this paper is to explore the quality of Dynamic Open Government Data for the development of data intelligence applications. Towards this end, we study a single case about the traffic data provided by the official Greek OGD portal. The portal involves the use of an Application Programming Interface (API), which is essential for the effective dissemination of dynamic data. Our research approach involves the exploration and the evaluation of the provided data with regards to missing values and anomalies. We anticipate that this paper will contribute to the identification of organisational and technical challenges that hamper the effective dissemination of dynamic OGD.

CCS Concepts: • **Applied computing** → **E-government**; • **Information systems** → **Data management systems**.

Additional Key Words and Phrases: Open government data, dynamic data, traffic data, sensor data, data quality

ACM Reference Format:

Areti Karamanou, Petros Brimos, Evangelos Kalampokis, and Konstantinos Tarabanis. 2022. Exploring the Quality of Dynamic Open Government Data for Developing Data Intelligence Applications: The Case of Attica Traffic Data. In *26th Pan-Hellenic Conference on Informatics (PCI 2022), November 25–27, 2022, Athens, Greece.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3575879.3575974

1 INTRODUCTION

Open Government Data (OGD) is a political priority the last decade in many countries in order to harness multifaceted benefits including enhancing evidence-based policy making and stimulating economic growth. OGD are expected to improve decision-making processes [18], to stimulate economic growth and innovation [23], and to provide opportunities

*All authors contributed equally to this research.

Manuscript submitted to ACM

for the development of more effective public services [7], including Integrated Public Services (IPS) [20]. Although the potential economic value of OGD is assessed in millions and billions [19], their full potential has not been exploited yet.

European Commission recently recognised dynamic data (including environmental, traffic, satellite, meteorological, and sensor generated data) as an important part of OGD presenting huge potential economic value [15]. It is indicative that the majority of the national OGD portals disseminate dynamic data [14]. Examples of this data include road traffic data, passenger ticket validation data, commercial shipping traffic etc. The immediate availability and regular updates of these data are crucial for the creation of added value data-driven services and applications [15].

Data quality is an important driver for the performance of OGD initiatives [6] and also for citizens trust in OGD [16]. This is particularly true in the case of dynamic data. In sensor data, although it is common to have errors in data, poor data quality render them useless and may lead to wrong decision-making results [21]. Data quality issues include the number, richness of content, and timeliness [22], the frequency of update of sensor data [14], as well as their insufficient granularity [11].

The objective of this paper is to explore the quality of Dynamic Open Government Data that could potentially facilitate the development of data intelligence applications. Towards this end, we focus on and study a single case, namely the traffic data of the Region of Attica that are provided through data.gov.gr, the official Greek OGD portal. This case was selected because it involves the use of an API that ensures the immediate availability and regular updates of the data.

Our research approach involves the exploration and the evaluation of the provided data regarding the existence of missing values and anomalies. Anomaly detection comprises both the identification of (a) anomalous flow-speed correlations and (b) deviations from the normal traffic pattern. These two present complementary views of the traffic data quality.

This paper is organised as follows. Section 2 presents the background knowledge required to understand the content of this work. Section 3 describes the approach used to explore the quality of sensor data. Section 4 provides the details of the case vignettes of this work. Section 5 detects the missing values in the sensor data, while Section 6 detects anomalies in sensor data using three methods and presents a case for the data of a specific sensor. Finally, Section 7 discusses the results of this study.

2 BACKGROUND

This section presents the background knowledge required to understand the content of this paper. Specifically, it describes Dynamic Open Government Data and data intelligence applications with traffic data.

2.1 Dynamic Open Government Data

Open Government Data (OGD) are a political priority the last decade in many countries in order to harness multifaceted benefits including enhancing evidence-based policy making and stimulating economic growth. OGD are data published by the public sector in open and reusable formats without restriction or charge for their use by society [10]. OGD have a huge potential. For example, the large volumes of OGD daily produced in the urban ecosystems can facilitate the creation of innovative products and data intelligence applications (e.g., [5, 9]) that monitor and analyze OGD in order to enhance the delivery of public value to the society and ensure a better quality of life [13].

Dynamic data (including environmental, traffic, satellite, meteorological and sensor generated data) have been recently recognised as an important part of OGD [15]. The immediate availability and regular updates of these data

are crucial for the creation of added value services and applications. Hence, dynamic data should be made available immediately after collection via an application programming interface (API).

2.2 Data Intelligence Applications with Traffic Data

Data intelligence applications exploiting traffic data involve predicting various traffic states such as traffic flow, traffic speed, and traffic demand [8]. Traffic flow is the number of vehicles passing a single spatial point (e.g., a road segment or a traffic sensor) in a specific period of time (e.g., number of vehicles per hour). Traffic speed is defined as the average speed of all vehicles passing a specific spatial point in a given time period and its prediction is useful for measuring the estimated time of arrival (ETA) in web mapping services such as Google Maps [4]. Traffic demand is the potential demand for travel and its prediction can be useful for taxi and ride – hailing services (such as Uber or taxi – call platforms) to schedule the allocation of their resources including drivers and vehicles in areas with increased demand.

Traffic forecasting methods can be categorized into baseline and deep learning methods. Baseline methods, such as traditional linear time series models (autoregressive models), do not model the spatial dependency, while deep learning and machine learning methods (e.g., neural networks) manage to model this dependency [17]. Traditional approaches generally manage to predict future values of time series in a stochastic way using past data. The recent development of deep learning in fields like image recognition, speech recognition, and natural language processing indicates that large amounts of data can be processed and trained through deep learning algorithms with application in the traffic forecasting domain. This application is ensured with the large volume of open data sets that smart cities obtain over the last decades.

3 RESEARCH APPROACH

According to our approach we explore a single case, namely the traffic data of the Region of Attica that are provided through data.gov.gr, the official Greek OGD portal. This case was selected because it involves the use of an API that ensures the immediate availability and regular updates of the data. We collect the data and study their quality based on the following dimensions.

3.1 Missing Values Detection and Imputation

Since the traffic data used in this work were collected by sensors, there is a chance that some observations may be missing due to various reasons, such as failures of sensors, network faults, and other issues. In order to find missing values in the measurements provided by the sensors, the number of observations that should be available for all sensors each day should be calculated first as:

$$number\_of\_active\_sensors\_per\_day * 24\, hours$$

The number of missing values is then calculated by subtracting the number of observations from the sum of records that should be available for all days based on the above type.

In order to deploy an anomaly detection algorithm on a time series dataset, the missing values can be replaced with numerical values using an imputation method. Imputation methods for missing values are categorized to prediction methods, interpolation methods, and statistical learning methods. In our case, we implemented the simple linear interpolation method to fill the missing values from the sensor data. Linear interpolation method estimates the missing value by assuming a linear relationship between the missing and non – missing values. It estimates the missing value based on the values of the adjacent data points to the interpolated data point:

$$y = y_b + \frac{(y_a - y_b) * (x - x_b)}{(x_a - x_b)}$$

where (x, y) is the point with the missing value y and $(x_a, y_a), (x_b, y_b)$ are the adjacent points prior and after the missing value.

3.2 Anomaly Detection

In our research approach, anomaly detection comprises both the identification of (a) anomalous flow-speed correlations and (b) deviations from the normal traffic pattern. These two present complementary views of the traffic data quality. Towards this end, we employ the two following methods respectively.

3.2.1 Anomalous flow-speed correlation. In traffic data, the number of cars counted by a sensor and their average speed are strongly correlated. In particular, considering that each sensor measures data that pass from one or more lanes, the maximum number of vehicles that can pass in all lanes in one hour can be calculated as [2]:

$$number\_of\_cars = \frac{average\_speed * 1000}{average\_vehicle\_length + \frac{average\_speed}{3.6}} * number\_of\_lanes$$

where *average_speed* is the average speed provided by the sensors measured in km per hour and *average_vehicle_length* is the average length of the different types of vehicles, the fraction *average_speed /3.6* represents the "safe driving distance" that should be kept between vehicles and is based on the vehicle speed, and *number_of_lanes* is the number of lanes in the road each sensor is positioned. The value of *average_vehicle_length* is set to 4. When the number of cars measured by a sensor in an hour is higher than this value, then the measurement is considered as an anomaly.

3.2.2 Seasonal – trend decomposition using Loess for anomaly detection. The seasonal – trend decomposition of periodic time series using Loess (STL) is a fundamental method for time series analysis, with many applications in anomaly detection and forecasting [3]. The robust STL algorithm performs seasonal – trend decomposition using the the statistical smoother "locally estimated scatterplot smoothing" - "Loess" (a generalization of the moving average technique) nd locally – weighted regression functions to decompose the time series. Specifically, STL considers the original time series as a composition of three components (additive model):

$$y_t = T_t + S_t + R_t$$

where y_t is the observed data at time t, T_t denotes the trend in time series, S_t is the seasonal component of the original time series, and R_t denotes the remainder component. The trend component shows a general pattern in time series on the long-term basis, the linear increasing (uptrend) or decreasing (downtrend). Furthermore, the seasonal component refers to the repeating patterns (periodic patterns) over time. Finally, the remaining variations in time series are the remainder component, also known as the noise. The remainder component is calculated by subtracting the trend and seasonal component from the original series. Remainder curve indicates the existence of noise present in the data.

STL decomposition is very useful for anomaly detection in time series by analyzing the residual curve of the STL output time series. For that reason, after the decomposition procedure, the remainder curve is divided into an area of normal data points and an area of outliers or anomalies. The limits of these areas in the remainder curve can be defined by various methods, such as the InterQuartile Range (IQR) method or the empirical rule for normal distribution [1].

4 CASE VIGNETTES

4.1 The official Greek Open Government Data portal

Data.gov.gr is the official Greek data portal for Open Government Data. The latest version of the data portal was released in 2020 and provides access to data published by the central government, local authorities, or other Greek public bodies classified in ten thematic areas including environment, economy, and transportation.

The major update and innovation of the latest version of the Greek OGD portal was the introduction of an Application Programming Interface (API) that enables accessing and retrieving the data through either a graphical interface or code. The API is freely provided and can be employed to develop various products and services including data intelligence applications. In order to use the API, users need to get a token by completing a registration process and providing personal information (i.e., name, email, and organization) as well as the reason for using the API.

The introduction of the API enables the timely provision of dynamic data that are frequently updated. The API can be used, for example, to retrieve datasets describing data related to a variety of transportation systems (e.g., road traffic for the Attica region, ticket validation of Attica's Urban Rail Transport, and route information and passenger counts of Greek shipping companies). The frequency of data update varies.

4.2 Traffic Data in the Region of Attica

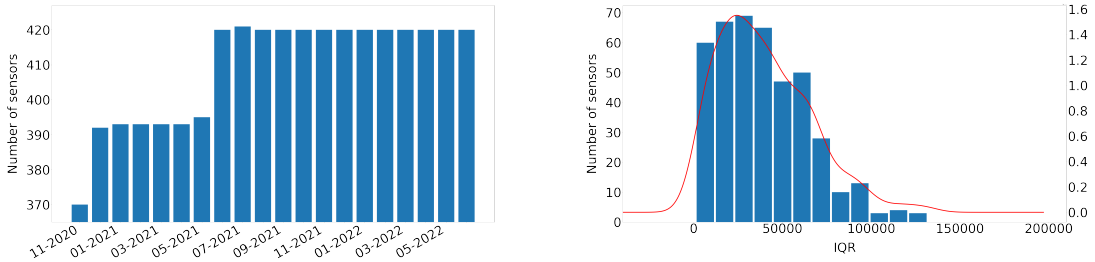
Traffic data for the Attica region in Greece are collected from traffic sensor nodes, which periodically transmit information regarding the number of vehicles in specific roads of Attica along with their speed. The data are hourly aggregated in order to avoid raising privacy issues. Data are hourly updated with only one hour delay.

We used the API provided by data.gov.gr and collected 4,230,819 records for a 22 month period, i.e., from 05/11/2020 to 31/06/2022. Each record includes (a) the unique identifier of the sensor (e.g., MS834), (b) the road in which the sensor is located along with (c) a detailed text description of its position, (d) the date and time of the measurement, (e) the absolute number of the cars detected by the sensor during the hour of measurement, and (f) their average speed in km per hour. The exact position of the sensor is a text description in Greek language and usually provides details including whether the sensor is located on a main or side road, or on an exit or entrance ramp, the direction of the road (e.g., direction to center), and the distance to main roads (e.g., "200 meters from Kifisias avenue").

The collected traffic data are coming from 425 sensors. We manually mapped the position of the sensors to latitude and longitude geographic coordinators in order to be able to present data into a map visualization. Specific position details are missing for one sensor (i.e., the sensor with identifier "MS339") making it impossible to find its exact coordinates.

According to the data, the sensors did not start operating at the same time and few of them stopped before the end of the period. Figure 1a presents the number of active sensors each month. Most sensors (370 or 87%) are active from the first month of the data (November 2020) and then the number of active sensors gradually increases to reach the 420 sensors on June of 2021. Specifically, during June 2021, 25 new sensors were introduced and, in the same month, 2 sensors stopped operating. A new sensor was also introduced in July 2021 resulting to 421 sensors and then a sensor stopped operating on August 2021 resulting again to 420 sensors. Finally, in December 2021 two new sensors were introduced and two stopped operating keeping the number of sensors stable to 420 until the end of June 2022. We excluded from this work the data related to sensors that stopped operating (namely sensors with ids 'MS136', 'MS137', 'MS858', 'MS1000', and 'MS1001'). We, finally, resulted in 4,228,021 observations.

We then calculated the interquartile range (IQR) of the counted cars measured by each sensor. Fig. 1b shows the right-skewed distribution of IQR, meaning that, probably, few of the sensors counted large number of cars.



(a) Number of sensors that are active each month.

(b) Distribution of the interquartile range (IQR).

Fig. 1. Active sensors per map and distribution of the interquartile range (IQR) of the number of cars measured by each sensor.

5 MISSING VALUES

In this Section, we search for observations that are missing from the traffic data based on two dimensions; (i) the time, and (ii) the sensors. In the first case we calculate the missing values per day, whilst in the second the missing values per sensor.

Considering all the measurements in the time frame between 05/11/2020 and 30/06/2022, and the day each one of the 420 sensors was introduced, the number of total potential observations would be 5,295,504. However, 1,067,483 observations (or 20.16 %) are missing . Fig. 2 presents the number of missing observations per day. The numbers of missing observations are increased until the end of May 2021. However, from June 2021 there is a significant decrease in the number of observations that are missing. Finally, the missing records seem to eliminate after January 2022.

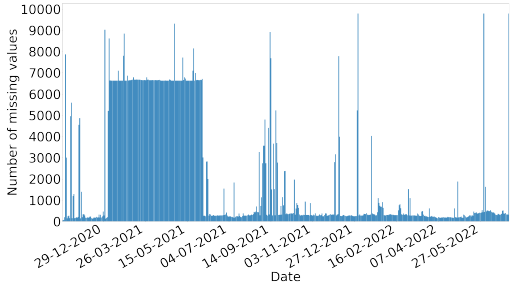
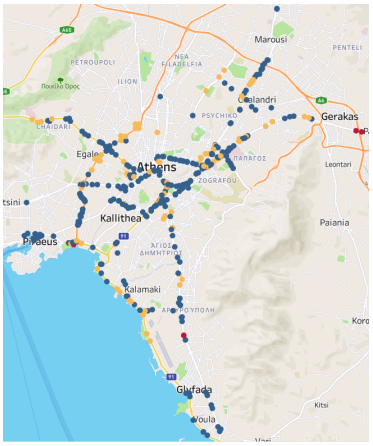


Fig. 2. Number of records that are missing per day.

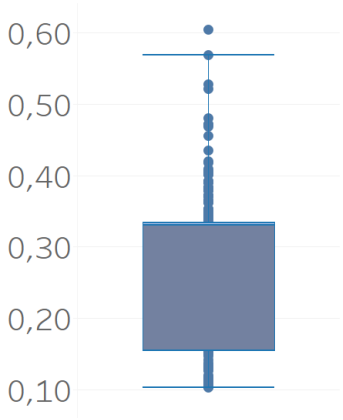
We also calculate the percent of missing records for each sensor from 05/11/2020 to 30/06/2022. Fig. 3 presents the street map with the 420 sensors and also the boxplot that presents the distribution of the missing values per sensor. The sensors are positioned in 93 main roads of the region of Attica. In Fig. 3a a mark is displayed over each sensor's longitude and latitude in the region of Attica. The colors of the marks indicate the percent of missing values of the sensor; red color marks represent sensors with higher percents of missing values, blue marks sensors with lower percents of missing values, and yellow marks sensors with intermediate percents of missing values.

In addition, the boxplot in Fig. 3b presents the distribution of the number of missing values per sensor. The median percent of missing values is 33.1% meaning that half of the sensors have less than or equal percents of missing values to the median, and half of the sensors have greater than or equal percents of missing values to it. The 50% of the sensors

have percent of missing values in the range 15.5% - 33.43% (interquartile range box). In addition, according to the whiskers of the boxplot (bottom 25% and top 25% of the data values, excluding outliers), the percent of missing values of each sensor may be as low as 10.3% and as high as 56.8%. In addition, based on our calculations, only 8 sensors have less than 10% missing values. Finally, only one sensor has percentage of missing values above 60%.



(a) View of the 424 sensors on a map. Each point in the map represents a sensor. Red color marks represent sensors with higher percents of missing values, blue marks sensors with low percents of missing values.



(b) Percent of missing values per sensor in a boxplot showing the lower (Q1) and upper (Q3) quartile, the median and mean values. Data falling outside the lower (Q1) - upper (Q3) quartile range are plotted as outliers of the data.

Fig. 3. Map of sensors and boxplot for the missing values per sensor.

6 ANOMALY DETECTION

6.1 Overview

In this Section we detect the anomalies in the traffic data based on two anomaly detection methods, namely (i) flow-speed correlation, and (ii) seasonal-trend decomposition with Loess. The time window we select to perform the above analyses is from 02/01/2022 to 22/06/2022. In this time window, the dataset includes 1,679,898 records with measurements produced by 420 sensors.

We first calculate the number and percentages of anomalies per sensor based on the flow-speed correlation filter described in section 3. In order to be able to calculate the number of vehicles that can pass in all lanes, we manually found the number of lanes that each sensor tracks and mapped them to the records. We discovered 1,230,928 records that count more vehicles than the number calculated by the filter (59.4% of total potential observations). We also calculate the number of anomalies per sensor. This number ranges from 0 to 3,853 anomalies. In addition, the mean number of detected anomalies per sensor is 2,937.8, while the median is 3,264 anomalies per sensor. There are only 15 sensors with less than 10% anomalies. Table 1 presents the descriptive statistics for the anomalies detected per sensor.

We also calculate the number of anomalies based on the STL method. For this method, identified missing values were imputed using the linear interpolation method as described in Section 3.2. According to Table 1, the mean number of anomalies detected per sensor (660.5 or 16.8%) is significantly lower than the number of flow-speed correlation

anomalies. STL anomalies per sensor ranges from 315 to 2024 anomalies. In addition, the median number of detected STL anomalies per sensor is 658 anomalies (or 16.11%) per sensor.

Table 1. Descriptive statistics considering the number of anomalies per sensor

	Flow-speed correlation (count)	Flow-speed correlation (mean)	STL (count)	STL (mean)
mean	2,937.8	71.1	660.5	16.8
standard deviation	821.7	19.9	174	4.26
min	0	0	315	7.71
first quartile	2,824	68.4	591	14.47
median	3,264	79	658	16.11
third quartile	3,416	82.75	711	17.4
max	3,853	93.3	2024	49.6

6.2 Anomaly detection - The case of sensor MS734

In this Section we present the details of the anomaly detection analysis of a single sensor based on the two methods, i.e., flow-speed correlation and Seasonal-Trend Decomposition using Loess.

For this case we selected sensor with id "MS734". We selected this sensor because it was listed as one of the sensors with the most correct values (0.19% of data were detected as anomalies) based on the flow-speed correlation filter in the previous Section. Sensor with id "MS734" is located at the regional unit of Piraeus within the Athens urban area. The sensor started operating at 05/11/2020 00:00. Based on the selected time window, the traffic dataset should include 4,128 records related to sensor MS734. Nevertheless there are 4,082 records with measurements of MS734, i.e., the dataset misses 46 or 0.011% of the measurements.

Table 2. Anomaly detection for the sensor with id "MS734"

total potential observations	actual observations	flow-speed anomalies	% flow-speed anomalies	STL anomalies	% STL anomalies
4,128	4,082	8	0.19%	571	13.9%

In addition, we apply the Seasonal-Trend Decomposition using Loess (STL) method for sensor "MS734" on the time window between 2022/01/02 and 2022/06/22. Figure 4 shows the decomposition of time series into trend, seasonal and residual for the selected sensor and time window. As mentioned in Section 3.2, after decomposing the original series we deploy anomaly detection on the residual curve of the output of STL.

The IQR method is applied on the residual curve in order to draw an upper and lower fence for outlier detection. The IQR method uses the following formula to detect anomalies, with values above the upper limit and below the low limit defined as outliers. We set the scalar multiplied with IQR to 3, after a set of experiments. Therefore, we noticed that setting low values to the scalar causes many observations to be considered as anomalies (Figure 5):

$$Upperlimit = Q_3 + 3 * IQR$$

$$Lowerlimit = Q_1 - 3 * IQR$$

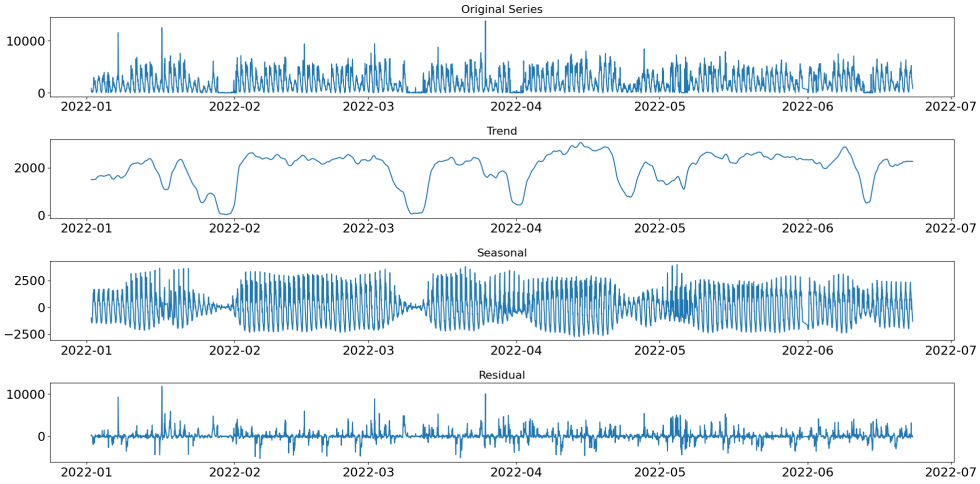


Fig. 4. Decomposition of time series (urban sensor "MS734" – 2022/01/02 00:00 – 2022/06/22 00:00).

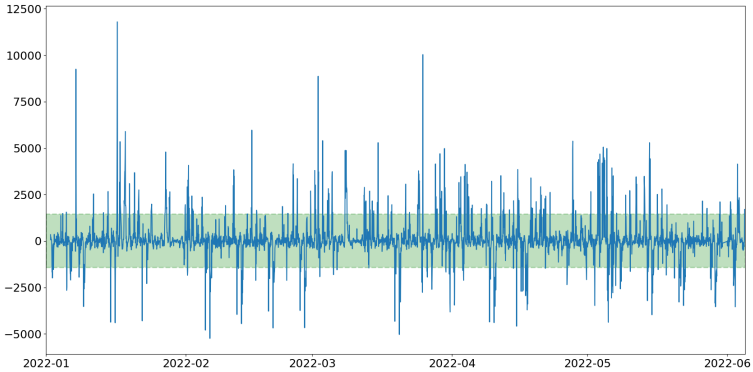


Fig. 5. Residual curve with the upper and lower limit of the IQR method for anomaly detection (urban sensor "MS734" – 2022/01/02 00:00 – 2022/06/22 00:00).

The residual curve (Figure 5) shows that the anomalies are detected both in positive and negative peaks on the remainder component of time series. Once anomalies have been detected on the remainder curve (also known as the noise), they are highlighted as red dots on the original series (Figure 6).

The STL decomposition applied to the aggregated data of sensor MS734 from January 2022 to June 2022 detects 571 anomalies (Figure 6) out of 4083 records. Table 2 shows the overall performance of sensor MS734 on traffic - flow filter and STL decomposition. The traffic - flow filter detects only 8 hours of anomalies, while the STL decomposition detects 571 anomalies. This is an observation that needs to be investigated further, finding the reasons behind this large amount of anomalies that STL detects. This could happen due to the fact that this particular sensor is located near traffic lights, thus the high peaks of counted cars are considered anomalous by the STL method which takes into consideration only seasonality and trend, while the traffic flow filter depends on fundamental attributes of a traffic state (such as the number of lanes and the correlation between speed and counted cars).

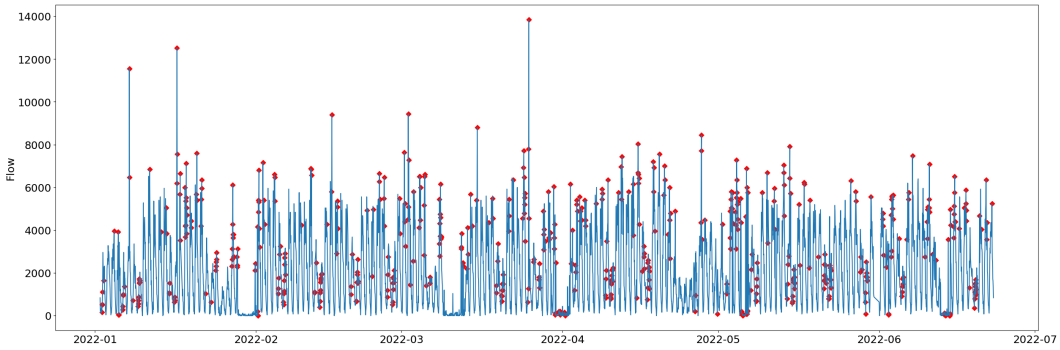


Fig. 6. Detected anomalies (red dots) of urban sensor "MS734" 2022/01/02 00:00 – 2022/06/22 00:00.

7 DISCUSSION AND CONCLUSION

Dynamic data are an important part of Open Government Data (OGD). Nevertheless, dynamic OGD are inherently vulnerable to quality errors hampering their involvement in the development of data intelligence applications. This paper explored the quality of Dynamic Open Government Data by focusing on a single case about traffic data of the Region of Attica. The data are provided by data.gov.gr, the official Greek OGD portal, through an API ensuring their immediate availability and regular update.

We found that, considering the time frame between 05/11/2020 and 30/06/2022, and the day each one of the 420 sensors that produced the data was introduced, 20.16% of the observations are missing. In addition, the 50% of the sensors have percent of missing values in the range 15.5% - 33.43% (interquartile range box). We also used two methods for detecting anomalies, namely the flow - speed correlation filter and STL decomposition. The flow-speed correlation filter found that the mean percent of anomalies per senor is 71.1%, while the same percent for STL is 16.8%. In addition, there are only 15 sensors with less than 10% anomalies based on the flow speed correlation filter.

These results suggest that further research is required regarding the organisational processes and technical approaches that are employed in the creation of the final data that are disseminated through the official OGD portal. Although it is well recognised in the literature that raw sensor data are vulnerable to data quality errors, governments need to carefully design the data pre-processing process in order to ensure that they, at least, do not increase and extend the raw data quality errors. For example, finer granularity of the disseminated data will enable end-users to more efficiently clean sensor data.

In the future, we plan to deploy additional anomaly detection algorithms like the Isolation Forest (iForest) algorithm on real time open data, to take advantage of unsupervised learning for unlabeled data. iForest is an unsupervised anomaly detection algorithm based on the hypothesis that outliers are always rare and few data points among the whole data - set (far from the center of normal clusters) [12]. Since the majority of real – world data - sets do not contain labeled anomalous data, unsupervised learning approaches are a suitable choice. Finally, we plan to use anomaly classification methods in order to classify the detected anomalies in anomalies that are (i) sensor errors and (ii) unusual traffic state (e.g., caused by accidents).

ACKNOWLEDGMENTS

This publication has been produced in the context of the EU H2020 Project inGOV which is co-funded by the European Commission under the Grant agreement ID: 962563.

REFERENCES

[1] Chiara Bachechi, Federica Rollo, and Laura Po. 2020. Real-Time Data Cleaning in Traffic Sensor Networks. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*. 1–8. https://doi.org/10.1109/AICCSA50499.2020.9316534

[2] Chiara Bachechi, Federica Rollo, and Laura Po. 2022. Detection and classification of sensor anomalies for simulating urban traffic scenarios. *Cluster Computing* 25, 4 (01 Aug 2022), 2793–2817. https://doi.org/10.1007/s10586-021-03445-7

[3] Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. 1990. STL: A seasonal-trend decomposition. *Journal of Official Statistics* 6, 1 (1990), 3–73.

[4] Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, et al. 2021. Eta prediction with graph neural networks in google maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3767–3776.

[5] Anne Gottfried, Caroline Hartmann, and Donald Yates. 2021. Mining open government data for business intelligence using data visualization: A two-industry case study. *Journal of Theoretical and Applied Electronic Commerce Research* 16, 4 (2021), 1042–1065.

[6] Mohammad Alamgir Hossain, Shams Rahman, Mohammed Quaddus, Elsie Hooi, and Abdus-Samad Temitope Olanrewaju. 2021. Factors Affecting Performance of Open Government Data Initiatives: A Multi-Method Approach Using Sem and FSQCA. *Journal of Organizational Computing and Electronic Commerce* 31 (2021), 300 – 319.

[7] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. Benefits, adoption barriers and myths of open data and open government. *Information systems management* 29, 4 (2012), 258–268.

[8] Weiwei Jiang and Jiayun Luo. 2022. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications* 207 (2022), 117921. https://doi.org/10.1016/j.eswa.2022.117921

[9] Evangelos Kalampokis, Areti Karamanou, and Konstantinos Tarabanis. 2021. Applying explainable artificial intelligence techniques on linked open government data. In *International Conference on Electronic Government*. Springer, 247–258.

[10] Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis. 2011. Open Government Data: A Stage Model. In *Electronic Government*, Marijn Janssen, Hans J. Scholl, Maria A. Wimmer, and Yao-hua Tan (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 235–246.

[11] Rob Kitchin and Sam Stehle. 2021. Can smart city data be used to create new official statistics? *Journal of Official Statistics* 37, 1 (2021), 121–147.

[12] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*. 413–422. https://doi.org/10.1109/ICDM.2008.17

[13] Fátima Trindade Neves, Miguel de Castro Neto, and Manuela Aparicio. 2020. The impacts of open data initiatives on smart cities: A framework for evaluation and monitoring. *Cities* 106 (2020), 102860. https://doi.org/10.1016/j.cities.2020.102860

[14] Anastasija Nikiforova. 2021. Smarter Open Government Data for Society 5.0: are your open data smart enough? *Sensors* 21, 15 (2021), 5204.

[15] European Parliament and European Council. 2019. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast). *Off. J. Eur. Union* 172 (2019), 56–83.

[16] Arie Purwanto, Anneke Zuiderwijk, and Marijn Janssen. 2020. Citizens' Trust in Open Government Data: A Quantitative Study about the Effects of Data Quality, System Quality and Service Quality. In *The 21st Annual International Conference on Digital Government Research* (Seoul, Republic of Korea) *(dg.o '20)*. Association for Computing Machinery, New York, NY, USA, 310–318. https://doi.org/10.1145/3396956.3396958

[17] João Rico, José Barateiro, and Arlindo Oliveira. 2021. Graph Neural Networks for Traffic Forecasting. *CoRR* abs/2104.13096 (2021). arXiv:2104.13096 https://arxiv.org/abs/2104.13096

[18] Erna Ruijer, Stephan Grimmelikhuijsen, and Albert Meijer. 2017. Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quarterly* 34, 1 (2017), 45–52. https://doi.org/10.1016/j.giq.2017.01.001 Open Innovation in the Public Sector.

[19] Anna Sołtysik-Piorunkiewicz and Iwona Zdonek. 2021. How society 5.0 and industry 4.0 ideas shape the open data performance expectancy. *Sustainability* 13, 2 (2021), 917.

[20] Efthimios Tambouris and Konstantinos Tarabanis. 2021. Towards Inclusive Integrated Public Service (IPS) Co-Creation and Provision. In *DG.O2021: The 22nd Annual International Conference on Digital Government Research* (Omaha, NE, USA) *(DG.O'21)*. Association for Computing Machinery, New York, NY, USA, 458–462. https://doi.org/10.1145/3463677.3463726

[21] Hui Yie Teh, Andreas W. Kempa-Liehr, and Kevin I-Kai Wang. 2020. Sensor data quality: a systematic review. *Journal of Big Data* 7, 1 (11 Feb 2020), 11. https://doi.org/10.1186/s40537-020-0285-1

[22] Dan Wu, Hao Xu, Wang Yongyi, and Huining Zhu. 2022. Quality of government health data in COVID-19: definition and testing of an open government health data quality evaluation framework. *Library Hi Tech* 40, 2 (01 Jan 2022), 516–534. https://doi.org/10.1108/LHT-04-2021-0126

[23] Zhenbin Yang, Sangwook Ha, Atreyi Kankanhalli, and Sungyong Um. 2022. Understanding the determinants of the intention to innovate with open government data among potential commercial innovators: a risk perspective. *Internet Research* ahead-of-print (2022). https://doi.org/10.1108/INTR-07-2021-0463