Evaluating the Health State of Urban Areas using Multi-source Heterogeneous Data

Dimitrios Tomaras and Vana Kalogeraki
Athens University of Economics and Business
Athens, Greece
{tomaras,vana}@aueb.gr

Nikolas Zygouras, Nikolaos Panagiotou and Dimitrios Gunopulos
University of Athens
Athens, Greece
{nzygouras,n.panagiotou,dg}@di.uoa.gr

***Abstract*—In recent years we are witnessing a growing interest in identifying various aspects affecting the quality of life in smart cities, such as traffic congestion and pollution levels, in order to provide services that enhance the public welfare. In smart cities, sensor infrastructures are deployed around the city combined with data analytics, to monitor and detect in real-time possible anomalies or events of interest. One major challenge that arise in smart-cities is to evaluate the health state of an urban city using heterogeneous multi-source urban data, *i.e.,* pollution and traffic data. Existing works in the literature are limited since they analyze a single source of data, either inferring the air quality or estimating traffic congestion. However, none of these works considers both data sources in concert for estimating the city's health state. In this work, we present "HELIoS" (HEalthy LIving Smart), a framework that combines multiple heterogeneous sources of data, *i.e.,* urban traffic and pollution data, to diagnose the health state of urban areas in a smart city. Our experimental evaluation provides valuable insights into identifying the health state of an urban area, and shows that our approach is both practical and efficient.**

I. Introduction

With the rapid increase in *urbanization*, cities are becoming more complex to live. It is estimated, that, by 2050 64.1% of the developing world and 85% of the developed world, will be urbanized, with more than 6 billion people living in cities. However, living in cities can be quite challenging for individuals, impacting their quality of life. Smart cities typically deploy sensor infrastructures combined with data analytics to monitor and detect in real-time possible anomalies or events of interest in the city (*i.e.* traffic accidents, air pollution, etc.). Sensor devices are becoming part of the smart city infrastructure and these can be either i) embedded in the city infrastructure (*e.g.*, SCATS sensors and CCTV cameras), ii) mobile (*e.g.*, smartphones or mobile apps) or iii) static (*e.g.* pollution monitoring stations) in order to provide services that enhance the public welfare and the citizens' quality of life. Such services may relate to traffic management, environmental monitoring, smart transportation, housekeeping information, etc. Smart city systems typically produce vast amounts of data during their operation, allowing us to study, understand and model several parameters of the city such as traffic patterns, human crowd behavior and dynamics, and air pollution diffusion across the city, as illustrated in Figure 1.

978-1-5386-4725-7/18/$31.00

Fig. 1. Typical example of pollution in urban areas.

Thus, digital technologies are changing the way in which smart cities are planned and monitored. Probably the most challenging question is how to evaluate the health state of an urban area on both spatial and temporal dimensions considering several factors such as the traffic conditions and air pollution across the city [1]. Monitoring air pollution emissions and concentrations and delivering this information to the citizens indicating the city's health state can have significant benefit on human habits and in the way people choose to live, behave and interact within their city environment. For instance, children, old aged or people performing physical exercises, such as bicycling or jogging, would benefit from such information for example to avoid leaving their homes or going near congested areas, where the pollution levels are high or citizens can be encouraged to use alternative transport solutions and drivers can avoid traffic congested areas that may result in increased levels of stress and aggression[2]. Therefore, identifying and extracting the city's health state using heterogeneous types of urban data plays a significant role in enhancing the public welfare.

Prior studies have analyzed the city living problem from various perspectives. For instance, in [3], the authors studied the optimal location for a new retail store using geographic criteria, where features are formulated according to the types and density of nearby places, and user mobility criteria, which includes transitions between venues or the incoming flow of mobile users from distant areas. There has been work focusing on inferring traffic information based on either crowdsourcing schemes [4], [5] or loop sensor networks [6], [7], [8], and on estimating the environmental footprint of a city using urban data[9]. However, these studies are limited as they focus on

individual aspects of the city living problem (*i.e.,* traffic or air pollution) and do not consider multiple types of urban data, which is the focus of this paper.

In this paper, we aim at evaluating the health state of an urban city using heterogeneous multi-source urban data, *i.e.,* pollution and traffic data. We have developed this work, in the context of the VaVeL project[1] that is currently *deployed* in the City of Dublin and aims at monitoring diverse data coming from city-wide infrastructures and recognize in *real-time* abnormal events of interest such as traffic conditions and the air pollution levels. In the VaVeL system, traffic data are received from various voluminous sources, including cameras CCTV, static loop sensors and bus sensors that measure the traffic flow, the data are possibly *noisy*, with *missing values* and *measurement errors*. In Dublin the air quality is measured city-wide through 6 stations deployed in the city center and its suburban areas to provide air pollution information to its citizens. Air quality does not stand still, it changes hourly, with large variations in quality even from neighborhood to neighborhood. Thus, evaluating the health state of the city using air pollution and traffic data is a challenging task: (a) although urban data are available to smart city authorities, it is very difficult for human operators to monitor this vast amount of information, (b) traffic data are available at significantly higher resolution compared to pollution data, and (c) identifying hidden connections between these two heterogeneous sources of urban data on both spatial and temporal is not a trivial task, since they may differ in sampling frequencies. Therefore, a more systematic and comprehensive analysis is required to understand and model how pollution emissions in the city are diffused in correlation with the road traffic density as measured by sensors deployed city-wide in order to efficiently diagnose the city's health state.

Contributions. In this work, we propose "HELIoS", a framework that identifies the city's health state by combining heterogeneous sources of urban data. Our contributions are summarized as follows:

- We propose "HELIoS", a scheme that aims at identifying the city's health state using traffic and pollution data.
- By conducting an extensive analysis on both data sources, we illustrate the hidden connections between the different types of data.
- We show how these different sources of data can be appropriately combined to build a regression model that diagnoses the health level of city areas.

The rest of the paper is structured as follows: In section 2 we motivate our problem and present our System Model. In section 3 we describe our approach. In section 4 we present our experimental evaluation. Section 5 describes related work and finally, section 6, concludes our paper with lessons learnt from this work.

[1]http://www.vavel-project.eu/

II. Preliminaries

In this section we first provide a brief motivation of our approach and then describe our system model.

A. Motivation

There has been extensive work in the literature studying how traffic congestion affects human health and well-being [2], [10], [11]. Their findings illustrate that dealing with heavy traffic does not just make people late for work or dinner; living with constant traffic congestion also has negative consequences on their health. Even healthy people can experience health impacts from polluted air including respiratory irritation or breathing difficulties during exercise or outdoor activities[12]. Therefore, it is important to study both factors to diagnose the city's health state.

Single-Source Data. Current body of research is focusing on analyzing data from single source feeds such as traffic flows [13], [14], [15]. These approaches rely on capturing mobility traces in order to estimate the traffic congestion and its location and extend in a smart-city area. For instance, geo-tagged social media data[16] is able to provide an indication about the area in which a traffic event occurs, but is not adequate to provide information regarding its environmental effect. Furthermore, detecting outliers in traffic flows[17] can reveal the area of a possible traffic event, but it says nothing about the degree at which the pollution affects the human health. In our previous work[18], we have focused on the traffic monitoring problem, and in this work, we extend our approach with analyzing multiple data sources for inferring the city health levels.

Multiple Sources of Data. Recent works in the literature utilize heterogeneous sources of data to infer human mobility [19], [20]. For instance, combining GPS probe data and tweets as in [20] may help in identifying a traffic congestion event. Furthermore, combining multiple sources of data, such as transit and cellphone data as in [19], it is possible to infer traffic congested areas. However, both studies do not provide any insights into inferring environmental pollution and as a result the city's health level. Moreover, in [21], the authors aim at analyzing the air quality, but their approach is not able to identify the city's health, although it is indeed impacted by traffic congestion. It is clear that these approaches are limited and could be used only to understand one of the two factors, but not both factors and their interplay. Our thesis is, that, pollution information can be used in combination with traffic data to infer the city's health level and this is the focus of the work we present in this paper.

B. System model

In smart cities, heterogeneous types of sensors are employed to facilitate the collection and processing of information related to traffic conditions as well as environmental conditions around the city. The work presented in this paper is conducted in the context of the Dublin City. Dublin is a typical example of a smart city that utilizes sensors for monitoring the traffic state across the city and the levels of hazardous pollutants,

such the nitrogen dioxide (NO_2)[7]. In order to achieve that, a variety of heterogeneous data sources are exploited including the following: (i) SCATS sensors which are embedded on the road and monitor real-time traffic density, (ii) GPS traces from sensors embedded on buses, (iii) the LiveDrive radio where users can report traffic conditions, (iv) pedestrian counters and (v) CCTV cameras that display the traffic conditions in real-time. For instance, the VaVeL Project instruments the above diverse data feeds in order to facilitate traffic monitoring.

SCATS System: SCATS (Sydney Coordinated Adaptive Traffic System) is an innovative computerized traffic management system developed by Roads and Maritime Services (RMS) Australia. SCATS sensors are fixed magnetic sensors deployed on intersections to measure various road characteristics such as the traffic flow and the degree of saturation of roads' lanes. In Dublin city, each SCATS sensor transmits a new record with a sampling frequency of one minute. Each one of these records comprises the following information: $<$ timestamp t_l of the measurement, the sensor's unique ID_i, degree of saturation measured at sensor ID_i at timestamp t_l denoted as ds_{i,t_l}, and traffic flow measurement at ID_i sensor at timestamp t_l denoted as $f_{i,t_l} >$.

At the moment, in Dublin city, there are approximately 557 SCATS controlled intersections and 3402 different SCATS sensors deployed throughout the road network. Each SCATS sensor monitors the traffic condition of a specific lane of a junction. The geospatial coordinates of SCATS sensors are presented in Figure 2. The degree of saturation illustrates how much a road's lane is utilized, while the traffic flow measures the vehicles' volume divided by the highest volume that has been measured in a sliding window of a week[2]. Both metrics are essential for understanding the specific conditions of the road network at any time. More formally, the degree of saturation ds_{i,t_l} is defined as:

$$ds_{i,t_l} = \frac{green(t_l) - (T_{total} - q(t_l) \cdot sp)}{green(t_l)} \tag{1}$$

where $green(t_l)$ denotes the duration of the green traffic light, T_{total} is the total time where no vehicle passes the sensor, $q(t_l)$ the time between vehicles while the sensor is discharging and sp the number of spaces between cars. The traffic flow f_{i,t_l} is defined as:

$$f_{i,t_l} = ds_{i,t_l} \cdot green(t_l) \cdot \frac{veh}{3600} \tag{2}$$

where $\frac{veh}{3600}$ is the number of vehicles per second at maximum flow. In this work we focus on the measurement of the degree of saturation, denoted as ds_{i,t_l}, as it is more reliable and informative than the traffic flow. The degree of saturation captures better the congestion of a road since, for instance, the traffic flow may have a low value almost equal to zero, but the road is congested because of a traffic event.

[2]http://dublinked.com/datastore/datasets/dataset-274.php

Pollution Monitoring: In smart-cities the pollution levels are measured over time from six monitoring stations installed around the city that show results for hourly Pollutants Nitrogen Dioxide (NO2), Nitrogen Monoxide (NO) and Particulate Matter (PM). Each station is characterized by the following: $< (lat_j, lon_j), A_R >$, where (lat_j, lon_j) correspond to its geospatial coordinates of the station, and A_R is the area of coverage in a radius R around the geospatial coordinates of the station. In Figure 3 we illustrate the pollution monitoring stations in Dublin City. To estimate the pollution levels we use the pollution concentration metric; this is a value that denotes the concentration of a specific pollutant in the air. In Dublin, the pollution concentration measurements are generated on a per-hour sampling basis and are appropriately stored by the Dublin City Council. Moreover, this concentration value can be described using models that incorporate several parameters such as the emission factors of the pollutant, the earth topology etc. It has been shown[22] that the pollution concentration of an area A, $C_{A,t}$, at a specific time t can be described using a Gaussian plume model. More formally,

$$C_{A,t} = \frac{Q(t)}{4\pi Kx} e^{\frac{-y^2 u}{4Kx}} [\frac{1}{e^{\frac{(z-H)^2 u}{4Kx}}} + \frac{1}{e^{\frac{(z+H)^2 u}{4Kx}}}], x, y, z \in A \tag{3}$$

Health Level Index: Our objective is to diagnose the health level of a city area using a metric that captures both environmental and traffic conditions. Since both sources are different, we need to define a normalized metric that fuses information from both sources in order to identify the health level. Therefore, we define the health level index of a city area A at time t_l as the product form of the level of pollution of the area multiplied by the level of congestion of the area. Specifically, we normalize the pollution concentration divided by the value annotating hazardous health concerns C_{max} and we multiply it with the normalized average value of the degree of saturation for all sensors in area A (we divide by the degree of saturation value ds_A^{max} that annotates the occurrence of a traffic jam (equal to 160 for the Dublin case)). More formally, the health level index of an area A at time t_l is defined as follows:

$$HLI_{A,t_l} = (\frac{C_{A,t_l}}{C_{max}}) \cdot (\frac{\sum_{\forall i \in A} ds_{i,t_l}}{|A|} \cdot \frac{1}{ds_A^{max}}) \tag{4}$$

Problem Definition: Given a set of labels $\mathcal{B}$ that characterize the health levels of a given area A_R, and the two heterogeneous data sources $\mathcal{T}$ (for traffic) and $\mathcal{P}$ (for pollution), the problem is to accurately classify the city area into health levels, specifically, learn a function $\mathcal{M}$ that decides the health level of the city area A_R. More formally,

$$\mathcal{M}(\mathcal{T}, \mathcal{P}, A_R) \longrightarrow \mathcal{B} \tag{5}$$

III. HELIOS APPROACH

In this section, we present our approach to solve the problem. Our goal is twofold: first, we aim at identifying the level of traffic and air pollution across a city, and then build a regression model to identify the city's health state.

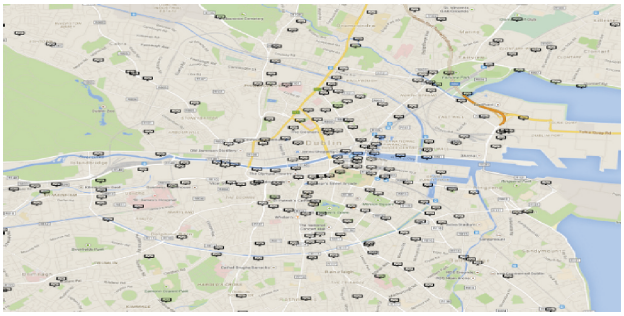


Fig. 2. The SCATS sensors scattered in Dublin's city center.

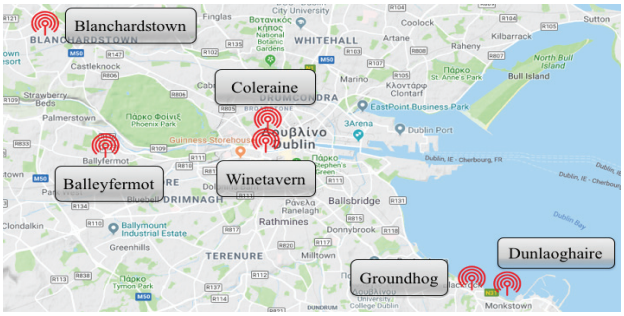
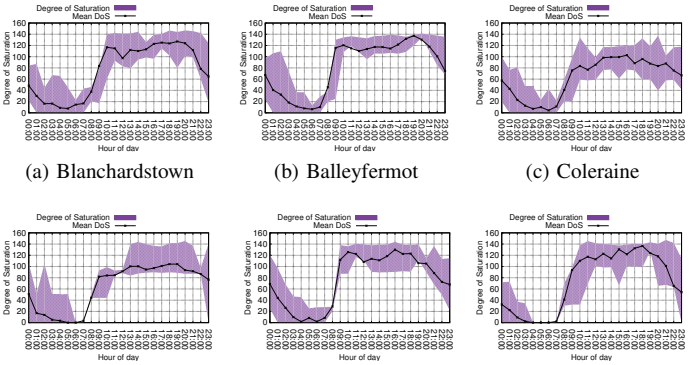


Fig. 3. Pollution sensors locations at the city of Dublin.



(a) Blanchardstown (b) Balleyfermot (c) Coleraine

(d) Winetavern (e) Groundhog (f) Dunlaoghaire

Fig. 4. Hourly Degree of Saturation (DoS) on typical days for districts of Dublin City

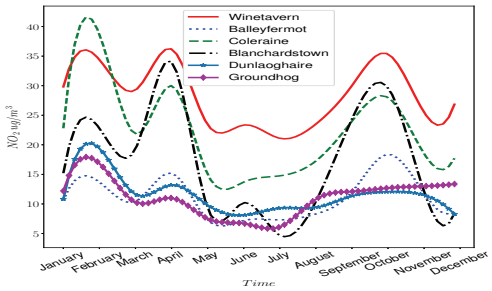


Fig. 5. Monthly median of pollution emissions for the available pollution sensors in the city of Dublin.

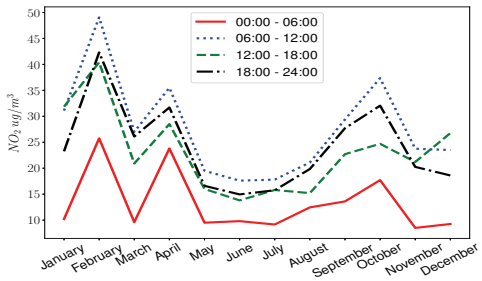


Fig. 6. Monthly pollution levels for Coleraine area for different hours of day.

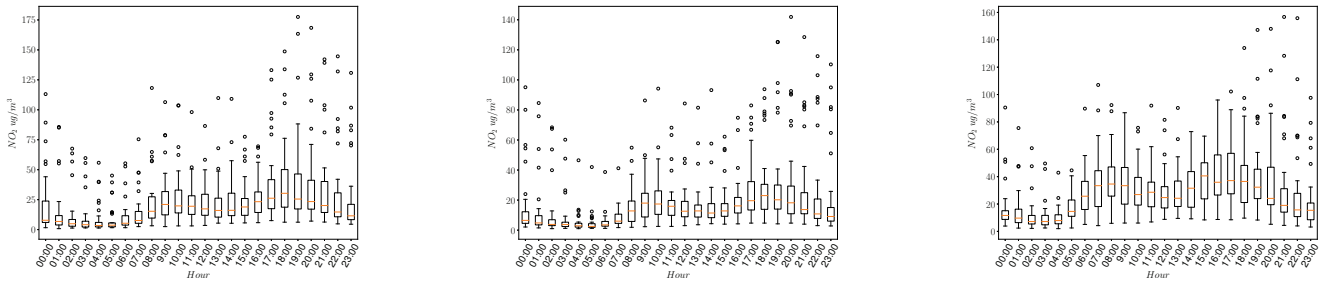
A. SCATS data analysis

In this section, we discuss our findings from the analysis of SCATS data. We first present the traffic density in Dublin City. In Figures 4a, 4b, 4c, 4d, 4e and 4f we illustrate the degree of saturation on typical weekdays on a 24hour basis for various districts around the city. We have chosen the degree of saturation as the appropriate metric to illustrate the traffic conditions, since it provides more meaningful information regarding the traffic state of the road network. For the purpose of the analysis, we draw the average degree of saturation for all sensors within a radius $R = 0.5km$ from the pollution monitoring station of each district on a per-hour sampling basis, along with the minimum and the max value observed as an envelope. Our goal was to identify how the degree of saturation varies over time and whether this can be described using a known distribution. Despite the fact that the different districts present different hourly mean and variance, we observe that the degree of saturation measurements can be approximated on an hourly basis using known distributions, such as the Gaussian Distribution. Thus, we may conclude that on typical days, the degree of saturation is expected to follow a known distribution and therefore it can be calculated using the statistical properties of this distribution.

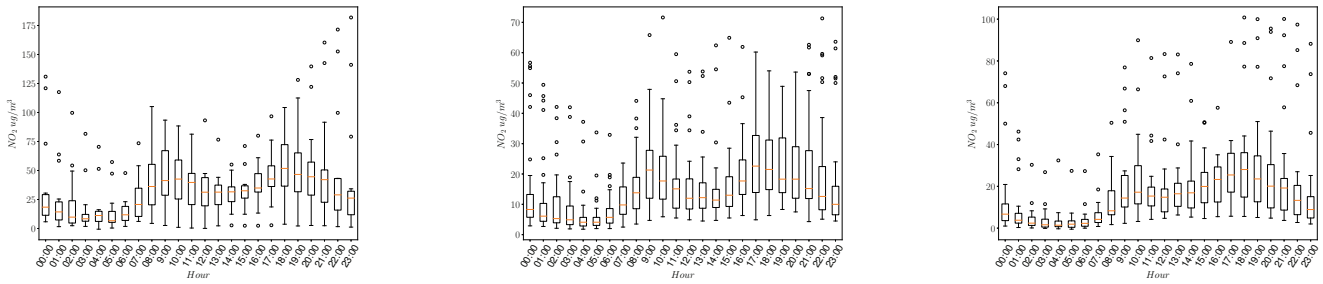
B. Pollution data analysis

We next focus on the analysis of the pollution data. We analyzed pollution data from the Dublin City for six different city districts (the locations of the monitoring stations are shown in Figure 3).

More specifically, in Figure 5, we illustrate the Nitrogen Dioxide (NO2) values as they are measured over time at the pollution stations in Dublin. We observe there are different pollution levels across the different regions of the city. Larger NO_2 emissions are observed at *Winetavern* area, while *Balleyfermot* and *Dunlaoghaire* areas exhibit lower emissions. We also observe that seasonality is present in the NO_2 measurements. The emission levels during the summer months are smaller than those in the winter months. This is happening since the pollution levels are not only related to the number of vehicles that are moving in the city, but they are also related

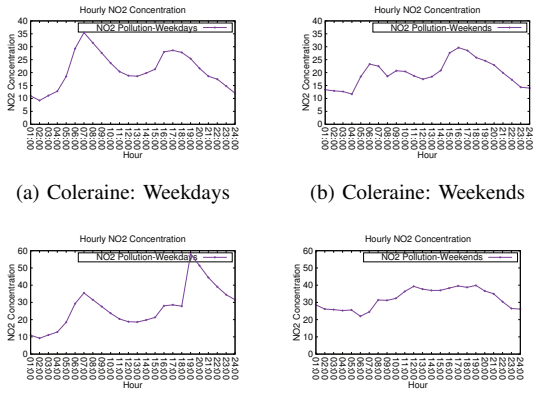


(a) Blanchardstown (b) Balleyfermot (c) Coleraine



(d) Winetavern (e) Groundhog (f) Dunlaoghaire

Fig. 7. Pollution Concentration over the day on typical days for the main districts of Dublin City



(a) Coleraine: Weekdays (b) Coleraine: Weekends

(c) Winetavern: Weekdays (d) Winetavern: Weekends

Fig. 8. Pollution in Dublin's city center districts.

to the temperature. Larger pollution rates are observed during the months with lower temperatures. In Figure 6 we present the median value for NO_2 emissions for the Coleraine Area of Dublin (a district located near the city center). As can be seen, the pollution levels during the night hours are considerably lower than the daily values. Finally from May till August the pollution level is considerably lower than the rest of the year. In Figures 7a, 7b, 7c, 7d, 7e and 7f, we illustrate the NO_2 emissions for different hours of day for the different districts on January. As we can observe the pollution varies significantly over the day. During the night hours (21:00 - 05:00) the pollution levels are significant lower than the daily values. This is explained due to the fact that fewer vehicles travel during the night and also fewer companies operate at night. Finally, we observe that the pollution measurements do not deviate considerably per hour.

Last, we chose two two city center districts (Winetavern and Coleraine) and analyzed their NO_2 concentrations. In Figures 8a, 8b, 8c and 8d, we draw the average hourly concentration of the pollutant for both weekdays and weekends. As we may observe, there is a significant difference between weekdays and weekends regarding the pollutant's concentration levels. This verifies our initial intuition that citizens rely on their cars for their daily commutes.

C. HELIoS Fusion-based Approach

In this section, we describe how we build our regression model for the NO_2 measurements using the SCATS data. We first describe how we aggregate the SCATS data in order to have similar shape with the environmental data. Following that we describe how we select the features for the regression technique. Finally, we present the regression models that we used.

SCATS Aggregation: The environmental measurements are reported every hour. On the other hand the SCATS data arrive in much higher frequency in the system. Every minute each SCATS sensor reports information regarding the traffic flow for a particular junction of Dublin. We aggregate the SCATS data measurements for each sensor computing the mean value of the degree of saturation metric every hour. In this way the SCATS data are transformed into a time series with 24 measurements per day, similarly to the pollution reports.

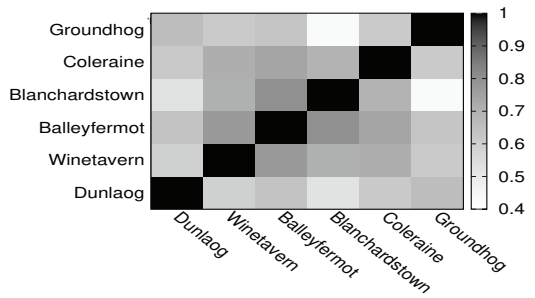


Fig. 9. Pollution Correlation among different districts of Dublin.

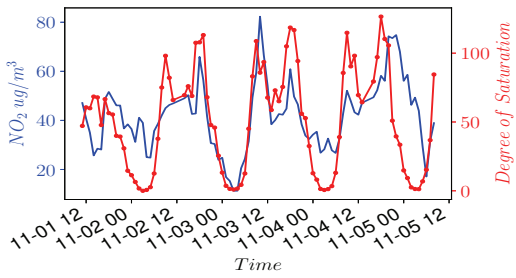
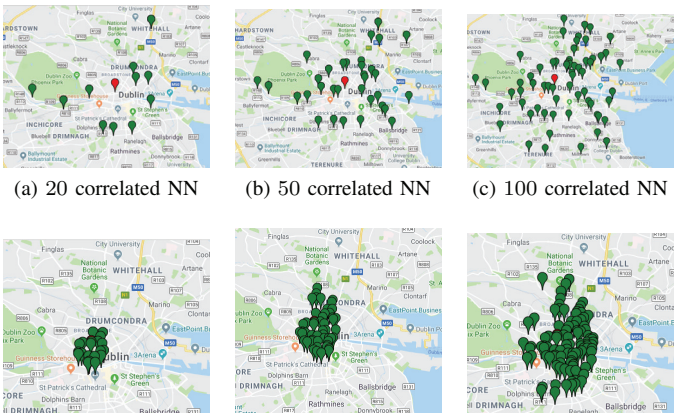


Fig. 10. NO_2 measurements (red) for the sensor located at Winetavern area and the degree of saturation measurements (blue) of the SCATS sensor with which it has the highest correlation.



(a) 20 correlated NN (b) 50 correlated NN (c) 100 correlated NN

(d) 20 spatial NN (e) 50 spatial NN (f) 100 spatial NN

Fig. 11. The 20, 50 and 100 SCATS sensors that are related either spatially or are correlated with the pollution station that is located at the Coleraine area. The red marker annotates the pollution monitoring station.

Feature selection: Here we describe how we select the SCATS data features that we use in order to estimate the NO_2 values for the different pollution stations in Dublin. The SCATS sensors are embedded at different locations around the city of Dublin (shown in Figure 2). In order to feed the regression models and preform the predictions for NO_2 values using the SCATS data we followed three different approaches that are described bellow:

- *Most Correlated:* We have observed that pollution concentration across different districts presents high correlation for neighboring areas, as illustrated in Figure 9. Hereby, we computed the correlation between each pollution station and each SCATS sensor and we selected the k most correlated SCATS sensors for each pollution station. For instance, Figure 10 illustrates the NO_2 measurements for the pollution station that is located at Winetavern area and the values of the SCATS detector with which it is mostly correlated. For instance, Figures 11a, 11b and 11c display the locations of the mostly correlated sensors with the pollution station that is located at the Coleraine area, in which, we can see that the locations of the sensors are distributed at different locations across the city.
- *Spatially Close:* In this case we select the k closest SCATS sensors with the location of each environmental station, considering the Euclidean distance between them. For instance, Figures 11d, 11e and 11f, demonstrate the locations of the SCATS sensors that are spatially close to the station located at the Coleraine area, the sensors are densely distributed near the area of the pollution station.
- *All the sensors:* In this scenario, we build the regression model for each pollution sensor using the values of all the SCATS sensors for the prediction. The regression model is responsible to learn the appropriate hyper-parameters/weights for each SCATS sensor.

Regression: In order to estimate the pollution data using the information available from the SCATS sensors we adapted a machine learning approach. The machine learning model uses historical data of observed pollution measurements and aims to estimate them using the information obtained from the SCATS sensors. During the training process, the goal is to identify the model parameters that result to the minimum estimation error. We used a variety of regression methods in order to identify the one that performs better for the pollution estimation task. For the different areas considered a distinct learner is trained. We utilized the following regression methods:

- Support Vector Regression (SVR): A robust regression variant of the Support Vector Machine classifier. Support Vector Regression is a non linear method if the appropriate kernel function is used.
- Random Forest: An ensemble that consists of multiple decision trees where each tree is trained using a different subset of the training set.
- Gaussian Process: The Gaussian process is a non-linear non parametric model and is an extension of the multivariate Gaussian distribution for infinite collection of real-valued variables.

Estimation of city's health level index: Having predicted the pollution concentration value from our regression model, we finally proceed on estimating the city's health level index for the particular area. We utilize the predicted value, which we divide by the maximum value that annotates hazardous pollution conditions, in order to derive the first term of the health level index equation (Equation 4). Then, given the actual

TABLE I
PREDICTION ACCURACY FOR THE DIFFERENT AREAS OVER THE VARIOUS AREAS USING **ALL** THE SCATS SENSORS.

	Gaussian Process			SVR			Random Forest		
Area	MAD	MAE	RMSE	MAD	MAE	RMSE	MAD	MAE	RMSE
Bally	32.43	33.20	36.05	12.98	12.89	15.00	13.37	13.97	16.27
Winetavern	42.60	43.57	45.83	10.06	11.91	15.24	10.54	11.25	13.76
Blanchard	43.73	46.58	53.02	21.48	22.85	27.14	18.86	21.49	25.70
Dunlaog	24.78	27.04	31.52	17.44	18.47	21.14	16.44	17.71	20.72
Coleraine	38.42	37.49	39.14	8.61	9.77	11.86	9.66	10.91	13.18
Average	36.39	37.58	41.11	14.11	15.18	18.08	**13.77**	15.07	17.93

TABLE II
PREDICTION ACCURACY FOR THE DIFFERENT AREAS OVER THE VARIOUS AREAS USING THE **SPATIALLY CLOSE** SCATS SENSORS.

	Gaussian Process			SVR			Random Forest		
Area	MAD	MAE	RMSE	MAD	MAE	RMSE	MAD	MAE	RMSE
Bally	32.43	33.20	36.05	12.99	12.87	14.97	11.85	12.85	15.02
Winetavern	42.60	43.57	45.83	10.06	11.91	15.24	9.52	11.19	14.13
Blanchard	43.73	46.58	53.02	21.46	22.79	27.09	17.90	19.96	24.22
Dunlaog	24.78	27.04	31.52	17.43	18.44	21.11	15.48	17.77	20.46
Coleraine	38.42	37.49	39.14	8.61	9.77	11.86	9.70	10.66	12.88
Average	36.39	37.58	41.11	14.11	15.16	18.06	**12.89**	14.49	17.34

TABLE III
PREDICTION ACCURACY FOR THE DIFFERENT AREAS OVER THE VARIOUS AREAS USING THE **CORRELATED** SCATS SENSORS.

	Gaussian Process			SVR			Random Forest		
Area	MAD	MAE	RMSE	MAD	MAE	RMSE	MAD	MAE	RMSE
Bally	32.43	33.20	36.05	12.98	12.89	15.00	12.73	14.16	16.79
Winetavern	42.60	43.57	45.83	10.06	11.90	15.23	9.28	11.15	13.70
Blanchard	43.73	46.58	53.02	21.47	22.84	27.14	17.85	20.17	24.77
Dunlaog	24.78	27.04	31.52	17.42	18.46	21.14	16.82	18.40	21.06
Coleraine	38.42	37.49	39.14	8.61	9.76	11.86	9.10	10.57	12.69
Average	36.39	37.58	41.11	14.11	15.17	18.07	**13.16**	14.89	17.80

value of degree of saturation, as reported from the SCATS sensors, we estimate how much congested the road is. This is performed by dividing this value with the maximum degree of saturation value, as given by the city authorities (please note that degree of saturation values over this threshold annotate a congested road). Finally, we multiply both values, in order to derive the city's health level index, and therefore, find the appropriate label that characterizes the city's health level (safe, moderate, hazardous).

IV. EXPERIMENTAL EVALUATION

In this section we present our findings regarding the correlation between the different sources of data. We evaluated our approach with respect to different aspects of the problem.

A. Experimental Setup

Dataset. We use two heterogeneous sources of data in order to evaluate our proposed algorithm. The datasets provide information regarding the road traffic conditions in Dublin City and the pollution concentration of NO_2 in districts of the city (both city center and suburban areas).

SCATS data: For the evaluation purposes, we used the SCATS data available for the period of time which overlaps with the period of time of our pollution data. We utilized the degree of saturation as the appropriate metric to observe how much a road is utilized. The SCATS data used in this work, contain aggregated data on hourly basis for 3402 sensors installed across 557 junctions. The available data used for the evaluation cover the period from 01/11/2015 until 21/12/2015. The SCATS sensors that occasionally fail to provide measurements are not used for the evaluation.

Pollution data: For the evaluation purposes, we utilized pollution monitoring data that we derived from the Dublin City Council. The pollution data contain information about the concentration of NO_2 and NO pollutants for six different districts of the Dublin City. Each record of the dataset provides information about the date, the hour of day and the respective concentration of the pollutant near the monitoring station. The

available data used for the evaluation expand from 01/01/2015 up to 30/04/2017.

Evaluation Metrics. In order to evaluate the accuracy of the proposed work the decision was to use the metrics: (i) Mean Absolute Error (MAE), (ii) Root Mean Squared Error (RMSE) and (iii) Median Absolute Deviation (MAD). RMSE in comparison to MAE gives high weights to large errors while MAD on the other had is not affected by large errors. The evaluation was performed under 5-fold cross validation. Since the data are time-series we performed cross-validation without shuffling in order to preserve the initial order. For all the machine learning methods we used the available implementation from the Scikit-Learn[3] Python library.

B. Experimental Results

In Table I we describe the results obtained when using all the SCATS sensors available in the city in order to estimate the pollution for a specific area. As the results illustrate, in terms of MAD, the SVM and the Random Forest approaches result to the best performance. The average MAD over all areas considered is 13.77. The drawback of this approach is the fact that since all the sensors are used, the computational requirements for training the model and performing the estimations are increased.

The above problem is solved using only a subset of the sensors to perform the estimation. In table II we describe the results obtained when using the SCATS sensors installed on the Top-50 spatially close junctions. Under this scenario, the best performing regression method is the Random Forest and the closest competitor is the Support Vector Machine. The average MAD over all the areas considered is 12.89 when using the Random Forest. This is a $\approx 6\%$ reduction on the estimation error in comparison to using all the SCATS sensors.

In Table III we present the results when using the Top-50 most correlated SCATS sensors for an area. According to the results, the best performance is achieved by the Random Forest. The average MAD over all the areas when using the Random Forest regression is 13.16. This fact suggests that the estimation accuracy is better than using all the SCATS sensors but slightly worse in comparison to using the spatially close sensors.

According to the above results it is clear that using a carefully selected subset of the SCATS sensors, in order to estimate the pollution for an area, may lead to similar or even better accuracy in comparison to using all the available sensors. The above realization could be very useful in cases where only a small number of sensors is installed across the city. In addition, the smaller the number of sensors used for the estimation the fewer the computational requirements in order to train the model and make the predictions. Figure 12, illustrates the MAD achieved when using the spatially close sensors with the Random Forest regression for different number of spatial neighbors used. According to the figure, in most of the cases there is no benefit by using more distant neighbors for the estimation.

[3]http://scikit-learn.org version 0.18.1

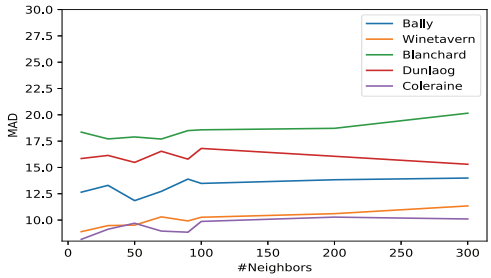


Fig. 12. The median absolute error (MAD) according to different number of spatial neighbors using the Random Forest regression. In most of cases, using a small number of neighbors results to similar or even better performance than using all the sensors.

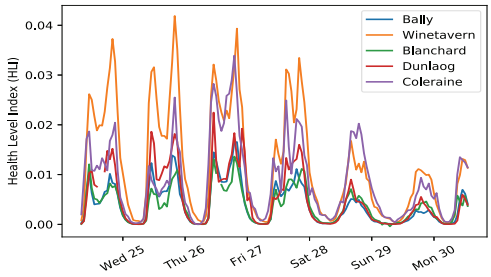


Fig. 13. The Health Level Index (HLI) for some days of November of 2015 for the various areas considered.

HELIoS Perfomance. In Figure 13 we illustrate the Health Level Index (HLI) as identified by HELIoS for several days of November 2015. We observe that during weekdays there is significantly higher health risk in comparison to weekends. We also observed that in the weekdays there are two spikes per day that correspond to the morning and evening rush hours while during night the HLI drops to near zero levels. Moreover, we noticed that during the weekends the two spikes pattern is not present. We reason this to the fact that during weekends the majority of the citizens do not drive to their work in mornings and back to home in the evenings. Therefore, we can safely conclude that "HELIoS" succeeds in diagnosing the health level of each area.

V. Related Work

Traffic Inference. In the literature, approaches that utilized multiple sources of data for inferring the traffic state have been proposed. In comparison with our previous work[23], in this paper, we focused on a totally different aspect, which is the estimation of the city's health levels using multiple sources of data. Moreover, the work of [20] utilizes multiple sources of data, such as tweets and GPS probe data to identify traffic congestion. However, the focus of their work is fundamentally different from the work presented in this paper and could not be applied to our setting. Authors of [24], introduce a macroscopic model regarding the estimation of traffic flow. However, their work focuses only on the traffic

estimation part and not in diagnosing the city's health levels. In contrast, in our work, we focus on a cross-domain fusion of heterogeneous types of data. The work of [4], incorporates a stacked autoencoder model to learn generic traffic flow features, but does not incorporates any pollution data so as to estimate the health levels of a city environment, whereas in our work, we build a model based on the correlation between traffic and pollution data. Several works over the last decades use loop detectors in order to estimate the traffic across the city. More specifically loop detectors were used in order to estimate the travel time of a given query path. [25], [26], [6] proposed techniques that estimate the vehicles' speed when crossing the loop detectors and then they converted the speed in the into the road segments' travel time.

Pollution inference. Authors of [27] present a prototype client-cloud system for pervasive and personal air-quality monitoring at low cost, which helps identify the concentration of particulate matters. However, their focus is totally different in comparison with the work presented in this paper, since, we aim at combining streams of heterogeneous data sources, such as pollution and traffic data. In [9], authors aim at deriving high resolution air pollution maps using mobile sensor nodes. However, in their work they do not incorporate the connection between the observed pollution data and actual traffic, as we do in our work. Finally, the work of [21] propose a semi-supervised learning approach to infer information about urban air quality. However, their work is limited since they do not focus on identifying the city's health levels, but only estimating the air quality.

VI. Conclusions

In this work, we proposed "HELIoS", a framework that identifies the city's health state by combining multiple heterogeneous sources of data. In our work we have made several findings:

- We conducted an extensive analysis on heterogeneous data sources and illustrated the relationship between the different types of data.
- We show how these different sources of data can be appropriately combined to build a regression model that diagnoses the health level of city areas and extract meaningful insights related to the health risks of an urban city.
- We illustrated that "HELIoS" is practical and efficient for recognizing health risks in smart cities.

Acknowledgment

This research has been financed by the European Union through the FP7 ERC IDEAS 308019 NGHCS project, the Horizon2020 688380 VaVeL project and a Google 2017 Faculty Research Award.

References

[1] M. Su, Z. Yang, and B. Chen, "Set pair analysis for urban ecosystem health assessment," *Communications in Nonlinear Science and Numerical Simulation*, vol. 14, no. 4, pp. 1773–1780, 2009.

[2] D. A. Hennessy and D. L. Wiesenthal, "Traffic congestion, driver stress, and driver aggression," *Aggressive behavior*, vol. 25, no. 6, pp. 409–423, 1999.

[3] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo, "Geo-spotting: mining online location-based services for optimal retail store placement," in *KDD*. ACM, 2013, pp. 793–801.

[4] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.

[5] I. Boutsis and V. Kalogeraki, "Crowdalert: a mobile app for event reporting and user alerting in real-time," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 261–264.

[6] J. Rice and E. Van Zwet, "A simple and effective method for predicting travel times on freeways," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 3, pp. 200–207, 2004.

[7] D. Kinane and et al., "Intelligent synthesis and real-time response using massive streaming of heterogeneous data (insight) and its anticipated effect on intelligent transport systems (its) in dublin city, ireland," in *ITS*, Dresden, Germany, November 2014.

[8] N. Zygouras and et al., "Towards detection of faulty traffic sensors in real-time." in *MUD@ICML*, 2015, pp. 53–62.

[9] D. Hasenfratz and et. al, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervasive and Mobile Computing*, vol. 16, pp. 268–285, 2015.

[10] J. I. Levy, J. J. Buonocore, and K. Von Stackelberg, "Evaluation of the public health impacts of traffic congestion: a health risk assessment," *Environmental health*, vol. 9, no. 1, p. 65, 2010.

[11] G. W. Evans and S. Carrère, "Traffic congestion, perceived control, and psychophysiological stress among urban bus drivers." *Journal of Applied Psychology*, vol. 76, no. 5, p. 658, 1991.

[12] K. Zhang and S. Batterman, "Air pollution and health risks due to vehicle traffic," *Science of the total Environment*, vol. 450, pp. 307–316, 2013.

[13] C. Zhang and et. al, "Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning," in *WWW*. International World Wide Web Conferences Steering Committee, 2017, pp. 361–370.

[14] W. Zhang, G. Qi, G. Pan, H. Lu, S. Li, and Z. Wu, "City-scale social event detection and evaluation with taxi traces," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 40, 2015.

[15] X. Zhou, A. V. Khezerlou, A. Liu, Z. Shafiq, and F. Zhang, "A traffic flow approach to early detection of gathering events," in *SIGSPATIAL*. ACM, 2016, p. 4.

[16] X. Liu, X. Kong, and Y. Li, "Collective traffic prediction with partially observed traffic history using location-based social media," in *CIKM*. ACM, 2016, pp. 2179–2184.

[17] J. Guo, W. Huang, and B. M. Williams, "Real time traffic flow outlier detection using short-term traffic conditional variance prediction," *Transportation Research Part C: Emerging Technologies*, vol. 50, pp. 160–172, 2015.

[18] N. Panagiotou and et al., "Insight: Dynamic traffic management using heterogeneous urban data," in *ECML*. Springer, 2016, pp. 22–26.

[19] D. Zhang and et. al, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *MobiCom*. ACM, 2014, pp. 201–212.

[20] S. Wang, L. He, L. Stenneth, S. Y. Philip, Z. Li, and Z. Huang, "Estimating urban traffic congestions with multi-sourced data," in *MDM*, vol. 1. IEEE, 2016, pp. 82–91.

[21] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *KDD*. ACM, 2013, pp. 1436–1444.

[22] J. M. Stockie, "The mathematics of atmospheric dispersion modeling," *Siam Review*, vol. 53, no. 2, pp. 349–372, 2011.

[23] D. Tomaras, V. Kalogeraki, N. Zygouras, and D. Gunopulos, "An efficient technique for event location identification using multiple sources of urban data," in *LocalRec Workshop*. ACM, 2017, p. 5.

[24] A. Nantes and et al., "Real-time traffic state estimation in urban corridors from heterogeneous data," *Transportation Research Part C: Emerging Technologies*, vol. 66, pp. 99–118, 2016.

[25] Z. Jia, C. Chen, B. Coifman, and P. Varaiya, "The pems algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors," in *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*. IEEE, 2001, pp. 536–541.

[26] K. F. Petty and et al., "Accurate estimation of travel times from single-loop detectors1," *Transportation Research Part A: Policy and Practice*, vol. 32, no. 1, pp. 1–17, 1998.

[27] Y. Cheng, X. Li, Z. Li, S. Jiang, Y. Li, J. Jia, and X. Jiang, "Aircloud: a cloud-based air-quality monitoring system for everyone," in *SenSys*. ACM, 2014, pp. 251–265.