Analysing Real World Data Streams with Spatio-Temporal Correlations: Entropy vs. Pearson Correlation

Maria Bermudez-Edo*

University of Granada, Granada, Spain

Payam Barnaghi^{a,}, Klaus Moessner^a*

^aUniversity of Surrey, Guildford, UK

Abstract

Smart Cities use different Internet of Things (IoT) data sources and rely on big data analytics to obtain information or extract actionable knowledge crucial for urban planners for efficiently use and plan the construction infrastructures. Big data analytics algorithms often consider the correlation of different patterns and various data types. However, the use of different techniques to measure the correlation with smart cities data and the exploitation of correlations to infer new knowledge are still open questions. This paper proposes a methodology to analyse data streams, based on spatio-temporal correlations using different correlation algorithms and provides a discussion on co-occurrence vs. causation. The proposed method is evaluated using traffic data collected from the road sensors in the city of Aarhus in Denmark.

Keywords:

Smart Cities, Internet of Things, Correlation, Entropy

1. Introduction

Globally more people live in urban areas than in rural areas. In 2050 it is expected that two-thirds of the population will live in urban areas [1]. This growth of the cities has given rise to the need for urban planning and to improve the infrastructure construction
5 [2]. In this context smart infrastructure, construction and building (SICB) systems will integrate basic infrastructures with smart sensing devices and intelligent applications that helps with the maintenance, monitoring and operation of the infrastructure systems. The volume of data produced on the Internet has increased exponentially in recent years, especially with the inclusion of sensory and Things' data. The data driven paradigms
10 look for transforming this massive information into actionable information and insights. This transformation is one of the main challenges of the Internet of Things (IoT). Special attention should be paid to the IoT in the cities and therefore urban computing for urban

*Corresponding author

Email address: mbe@ugr.es (Maria Bermudez-Edo)

planning is getting growing attention (see for example these two special issues [3] [4]).
 However, management and analysis of large volumes of data are still less developed
 15 than the capacity to collect data. Big Data analytics is particularly impacting the Civil
 Engineering domain and also in this field information systems are in a preliminary stage
 [5]. We face challenges in answering questions such as: How to use all this data? How
 to extract information and/or patterns and insights from it? One of the main starting
 20 points to analyse these big amounts of data is correlation detection that describe basic
 relationships between variables which is often used to derive a causal inference. One of
 the main goals of the correlation analysis is to reduce the information from large volumes
 of raw data into abstractions that describe the data [6]. Data analytics tools and big
 data algorithms heavily rely on correlations. Although different methods to analyse the
 correlations are available with differences in their results, previous research in the IoT
 25 and in particular in smart infrastructures, which is a prominent application domain of
 the IoT [7], has paid little attention to these differences. Furthermore, cities need to
 derive innovative solutions that can automatically infer urban dynamics and therefore
 to provide crucial information to urban planners [2] (e.g. [8]). For example, accurate
 estimation and prediction of urban travel times are essential for various applications in
 30 urban traffic operations and management [9].

This paper proposes an efficient method to derive spatio-temporal analysis of the
 data, using correlations, with Pearson and Entropy based methods and compares the
 results of both algorithms. Smart cities are an interesting field in IoT data management
 due to the multi-modal and multi-source nature of the data. IoT data can provide
 35 streams of information about cities and their citizens. One important part of the cities
 are the infrastructures and the construction. In smart construction there is not only
 an interest in using IoT at the construction phase, but it is also important to acquire
 information about the infrastructure usage and performance to optimise the operational
 efficiency, and this area will need a combination of efforts from different research fields
 40 [7]. We focus our research on inferring the urban dynamics to help in efficiently use the
 city infrastructures. In smart cities the lack of particular pieces of information or the
 faulty sources are common issues to deal with [10]. The efforts taken by some cities to
 became smart have involved the deployment of large and distributed sensor networks.
 The sensory data of neighbouring locations could be highly correlated. By effectively
 45 exploiting these spatial correlation it is possible to derive information from the data, or
 to infer the missing information.

The analysis of smart cities data is an emerging field of research and only a few works
 have analysed the correlations between different observation and measurement data in
 the context of cities (e.g. [11], [12]). All these works have utilised Pearson correlations in
 50 their analysis. However, it is well-known that Pearson correlation has some drawbacks.
 Pearson correlation fails to detect the dependency between two or more variables, when
 the data has specific distributions, such as non-linear distributions. In the field of statis-
 tics some alternatives are available such as distance correlation, mutual information or
 correlation ratio. Although none of the latter solutions are completely accurate, these
 55 alternatives could give a better hint of the dependency of data or could complement
 the Pearson correlation. Therefore, before describing our proposed solution we compare
 Pearson correlation with mutual information that is able to detect more general depen-
 dencies. We analyse both approaches, first with synthetic data and then with real data

60 taken from the smart open data Aarhus platform¹ in Denmark. In both cases we demonstrate that for certain data distributions mutual information could follow the dependency that Pearson correlation fails to follow, although Pearson has less computational cost.

The key contribution of this paper is to identify the correlations in multi-source data streams and to include the correlation metrics in a suitable analysis of real world data streams. We applied our analysis in smart infrastructures [7], which is a prominent field 65 using IoT sensors. In particular, this analysis gives a view of the movements inside the city networks; how different data related to people, traffic, electricity, etc. change around the city networks, road network, or water and electricity in pipe networks.

This paper highlights that conventional approaches of correlation analysis overlook the temporal component of the correlation. Common correlation analysis models often 70 rely on distance metrics between different patterns. However, when working with multi-modal data the distance measures do not always perform well in different situations. Following the spatial correlation will not be sufficient to model the behaviour of the data changes and their correlation in a city. We show a more generalised model by adding a temporal component to the spatial correlation analysis. This approach can be used 75 to model spatio-temporal correlations in different environments. We apply our proposed method to create a spatio-temporal traffic model in smart cities with IoT data coming from road sensors. The data is collected via bluetooth sensors that measure the number of cars in a road, attached to light poles (see Figure 1). These sensors are easier to deploy than traditional transportation sensors located on the ground, because they can 80 be installed without disrupting the traffic. They do not need high maintenance and include a GPS receiver that provides location data. The sensors are remotely configured and updated.² We use only the vehicle count variable, because we do not aim to create a complex model with several variables that in other cities should be difficult to replicate. Therefore our model can be extrapolated to other cities which only offer the vehicle count 85 information without other context information.

It is worth noting that correlations in the field of big data, and in the smart cities analytics, have been recently criticised^{3,4}. One of the biggest criticism is the assumption that correlation means causation, arguing that causation requires models and theories, not only correlations. However the classic causal science could not be effective when 90 dealing with complex problems and large datasets [13]. We argue that although smart cities data analytics strongly relies on correlation analysis, our approach is able to differentiate causation vs. correlation. Our approach uses the temporal correlation of data in a dynamic manner, looking for correlations in real time data and using the recent correlations to infer missing information or extract information, without questioning the 95 causation.

In summary, we first compare Pearson with entropy based correlations with synthetic and real data. The second, and main contribution is to model the urban mobility with a spatio-temporal analysis of urban data using both correlation techniques. Finally we discuss correlation vs. causation and explain how our analysis try to decouple them.

¹<http://www.smartaarhus.eu>

²<http://blipsystems.com/outdoor-sensor>

³<http://www.forbes.com/sites/gilpress/2013/04/19/big-data-news-roundup-correlation-vs-causation/>

⁴<http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html>

Figure 1: One of the sensors installed in a light poles. Source: Google Street Maps.

100 The remainder of the paper is organised as follows. Section 2 describes the related work. Section 3 gives a brief introduction to Pearson correlation and mutual information. Section 4 describes our proposal to model smart city data analysis based on spatio-temporal correlations. Section 5 describes the datasets and use-case scenarios that are used to evaluate Pearson correlation, mutual information and the proposed
 105 spatio-temporal model. Section 6 discusses causation and in section 7 we conclude the paper and discuss our future work in this domain.

2. Related Works

There are different types of sensors to measure smart cities parameters. In particular for vehicle detection, the traditional option is in-roadway sensors. Some of the most
 110 common in-roadway sensors are inductive-loop, presence-detecting magnetometers, and passage-detecting magnetometers. These sensors involve traffic disruption for its installation. Pavement deterioration, improper installation, and extreme weather conditions can degrade its operation. In the 1960s-1970s some large cities started to install sensors in poles, such as microwave radar or ultrasonic sensors. Recently new technologies have
 115 been developed that allow other types of pole sensors such as video processors and laser radars [14]. Currently other mobile technologies not specifically design for traffic monitoring are used in this field, such as in-vehicles location sensors [9] or unmanned aerial vehicles (UAV) [15]. UAV has specific challenges as the camera may rotate, shift and roll during video recording or shake due to wind fluctuations [16]. Furthermore it is difficult
 120 to have a continuous full picture of the city because only few UAVs are active and in a discontinuous manner.

Depending on the type of sensor used some preprocessing will be needed, such as image processing for vehicle detection, or basic operations to calculate average speed of vehicles detected. Video processing involves much complexity in the algorithms and time
 125 consuming [16].

Taking advantage of the sensor deployments in smart cities several recent studies applied big data to urban infrastructure operations [17], [5], [18], [19], [20]. A great

number of studies use sensors in taxis, mostly to optimize the taxi routes [21], [22], but also to study the mobility across the city [23] and its flaws to improve the urban planning [2]. Studies using mobile phones are also popular to obtain travel patterns [24], [20].

Several studies have analysed the city data changes (how different data related to people, traffic, electricity, etc. change around the city) mostly with the idea of better structuring the cities for future urban deployment [25], [26], [27], [28]. These studies do not include the temporal or dynamic component of the interactions. The lack of attention to potential spread of changes in the systems (e.g. unexpected events such as accidents, human reactions, or terrorist attacks at one location may affect other locations of the network) limits the possibility of using the data for accurate planning or for rapid reactions, such as evacuation or replacement of faulty services.

In vehicular traffic domain several works study the urban flows in the road infrastructures with micro and macroscopic fundamental diagrams [29], [30], [31], [32], [33] see [34] for a survey. However these diagrams frequently use contextual information different from the road sensors that sometimes cities do not provide. Furthermore these studies focus on the correlation between traffic density and vehicular speed either in one area or in the entire city. Our proposal is not tied to traffic data, it can be used with other types of data, and correlates the traffic in one sector (between two sensors) and other sectors with a temporal lag, providing not only a spatial correlation, but a spatio-temporal correlation analysis. Other works have studied the changes of traffic with different techniques (see [35] for a survey). For example, the historical traffic data and the status of traffic in neighbouring areas is usually used in providing predictions and analysis [36], [37], [38]. However a spatio-temporal analysis of the changes in traffic data in a city with individual road details has not been addressed to the best of the authors knowledge. This includes the impact that an event at one part of the city at a specific time will have on other parts of the city.

We propose an efficient method to derive spatio-temporal analysis of the data, using correlations. We use bluetooth sensors installed in light poles. Our proposed solution is computationally effective, compared to machine learning algorithms that require training and adaptations and are often limited in their applications.

3. Measures of Dependence Between Variables

In statistics there are several tools to measure the dependency of two or more variables. The first and the most common measure is Pearson correlation coefficient. This correlation measure is mostly sensitive to linear relationships between two or more variables. It also assumes a Gaussian distribution of the data. In smart city domain the data is often sporadic and the distribution of the data, especially in short term windows, is not often Gaussian; so the Gaussian methods are not always suitable for analysing and processing the multivariate smart city data. To overcome this constraint, other correlation coefficients have been developed to be more robust than the Pearson correlation. In particular we choose mutual information because it can measure non-monotonic relationships. Additionally the mutual information measure does not make any assumptions about the statistical distribution of the data. Hence, it is a non-parametric method. It also does not require the decomposition into modes, such those used in Principal Component Analysis (PCA) or in Independent Component Analysis (ICA). The latter is due to the fact that mutual information does not assume additivity of the original variables

[39]. In the following section we briefly describe the mathematical background of Pearson correlation and mutual information measures.

175 3.1. Pearson Correlation

Pearson's product-moment correlation coefficient (or Pearson coefficient) describes the proportion of the total variance in the observed data that can be explained by a linear model of the variables under study. It ranges from -1 to 1, with higher absolute values indicating better dependency between the variables.

180 The Pearson coefficient between two random variables x and y is defined as:

$$\rho_{x,y} = \frac{\sigma_{xy}^2}{\sqrt{\sigma_x^2 \sigma_y^2}} \quad (1)$$

Where:

- σ_x is the standard deviation of the variable x
- σ_x^2 is the variance of x
- σ_y is the standard deviation of the variable y
- 185 • σ_y^2 is the variance of y
- σ_{xy}^2 is the covariance of the variables x and y

There are several texts describing Pearson Correlation and demonstrating equation 1. See for example [40] [41] [42] [43] [44].

The coefficient of determination, $\rho_{x,y}^2$, gives the percentage of correlation between two variables. For example if $\rho_{x,y} = 0.5$, then $\rho_{x,y}^2 = 0.25$, means that 25% of the values of y are explained by the values of x .

One important factor in the measurement of dependency with Pearson correlation is the p value. The p value provides information about the probability that the data would be inconsistent with the hypothesis, that is, that the correlation is not significant, regardless of the value of $\rho_{x,y}^2$. It is generally assumed that a p value ≤ 0.05 implies that the correlation is significant and a p value > 0.05 implies that the correlation is not significant⁵.

Pearson correlation can be applied to measure the dependency between several variables by extrapolating the Pearson correlation coefficient to matrices of covariances and standard deviations.

200 When two variables are independent, the Pearson's correlation coefficient is 0. However, the inverse relation is not always true. There are some cases when the variables are dependent and Pearson's correlation does not detect the dependency. For example, Figure 2 shows a case in which Pearson's correlation does not detect the dependency. These type of cases appear because Pearson correlation mostly detects linear correlations between jointly normally distributed variables. If two variables are correlated but they do not follow a linear correlation, Pearson Coefficient could fail to detect the dependency.

⁵<http://www.jerrydallal.com/LHSP/p05.htm>

Figure 2: Examples of non-linear correlations that Pearson fails to detect.

3.2. Mutual Information

Mutual information is a quantity that explains how much one random variable tells us about another random variable. It measures the reduction in uncertainty about one random variable given knowledge of another. High mutual information indicates a large reduction in uncertainty; low mutual information indicates a small reduction; and zero mutual information between two random variables means the variables are independent. For example, a data stream (stream A) provides the number of vehicles in 500 meters of one road and separated by two kilometres another data stream (stream B) provides the number of vehicles in another 500 meters of the road. We can calculate the mutual information of streams A and B. If the mutual information is high it means that knowing the number of vehicles in stream A reduces the uncertainty of knowing the number of vehicles in stream B. We can infer the number of vehicles in stream B, if we know the number of vehicles in stream A as both streams are highly correlated. If the mutual information is zero, stream A does not give us any information about the number of vehicles in stream B. Mutual information is a generalisation of the correlation and is based on the concept of Entropy [45].

In order to define the mutual information, we need to introduce the Entropy concept shown as $H(X)$, and define the conditional entropy, $H(X|Y)$. Entropy is a measure of uncertainty of the information content in a random variable. Entropy is defined as:

$$H(X) = - \sum_x P_X(x) \log P_X(x) \quad (2)$$

Where:

- $P_X(x)$ is the probability distribution of x

Joint Entropy is a measure of uncertainty associated with a set of variables. Joint Entropy is defined as:

$$H(X, Y) = - \sum_x \sum_y P_{XY}(x, y) \log P_{XY}(x, y) \quad (3)$$

Where:

Figure 3: Relationship between the different Entropy concepts of two correlated signals.

- $P_{XY}(x, y)$ is the joint probability distribution of x , and y

The conditional Entropy, $H(X|Y)$, is the average uncertainty about X after observing a second random variable Y , and is defined as:

$$H(X|Y) = \sum_y P_Y(y) \left[- \sum_x P_{X|Y}(x|y) \log(P_{X|Y}(x|y)) \right] \quad (4)$$

Where:

- $P_{X|Y}(x|y) \equiv P_{XY}(x, y)/P_Y(y)$ is the conditional probability of x given y

There is a close relationship between all these concepts that represent the conjunction or disjunction of variables. The entropy $H(X)$ of the variable X , in Figure 3, (the left circle) is the uncertainty of the variable X . The entropy $H(Y)$ of the variable Y (right circle) is the uncertainty of the variable Y . Mutual information $I(X, Y)$ (the intersection part of Figure 3 in the centre), is the intersection of the uncertainty of the two variables. Mutual information is defined as the reduction in uncertainty about variable X after observing Y . This explanation can be easily derived by looking at Figure 3 and is expressed in equation 5. Values of the mutual information are always positive and are between 0 and the minimum value of the entropies of the variables: $I(X, Y) \in [0, \min(H(X), H(Y))]$. $I(X, Y) = 0$ indicate no correlation, values close to 0 indicate low correlation and values close to $\min(H(X), H(Y))$ indicate high correlation between X and Y .

$$I(X, Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y) \quad (5)$$

All these definitions can be extrapolated to multiple variables. Likewise, most results derived for discrete variables can be extend to continuous distributions, replacing sums by integrals. However, the replacement of sums by integrals hides a great deal of subtlety and for distributions that are not sufficiently smooth, the integral equations may not even work (although this is not normally the case). More details can be found in [46].

One constraint of mutual information is that it has a higher processing complexity than Pearson correlation. However, most of the current sensor nodes are equipped with processing capabilities and part of the correlation analysis can be done on the nodes that

collect the data [47]. Furthermore, some of these correlation analysis do not need to be executed in real time. Therefore, the constraint is application dependent and most of the times it can be overcome by a combination of leveraging the edge and cloud computing solutions.

260 4. Modeling City Movements based on Spatio-Temporal Correlations

Most of the data changes in the cities occur and disseminate over a network infrastructure. For example, water distribution runs over a network of pipes, social media data runs over a communication network, people or vehicles move over a network of streets and roads. These networks have an inherent spatial component, each node has
265 neighbouring nodes, and nodes that are linked through other nodes. However, the data changes have also a temporal component and elements such as people, vehicles or water have different patterns that cannot be simply modelled by the underlying network infrastructure topology. However, this movements affect the efficient use and future plans of the infrastructures.

We propose using spatio-temporal correlations to identify and represent correlating patterns. By using correlations we address not only the spatial topology, but also the actual patterns and data changes in a city. With this approach we can know how the status of one part of the network infrastructure can affect the status of other part in an elapsed period of time. For example, we can infer that a dense vehicle traffic in one point
270 of the city will affect another several minutes later. For this purpose we need data such as number of people in each street of the city or velocity of the vehicles in each road with relevant number of observations.

There are other approaches that try to cover the temporal distance between two signals, such as Dynamic Time Warping (DTW) and Frechet distance. However these
280 similarity distance measures are computationally expensive. For example, we performed an experiment using DTW between two signals and it took 4.7394 seconds, while the same experiments calculating both, Pearson correlation and mutual information, for 27 different time lags only last for 1.3850 seconds. Furthermore these similarity distance measures the similarity between two temporal sequences which may vary in speed, but
285 they do not provide a time lag where the correlation is maximum. Other approaches such as cross correlation provide also some correlation measures taking into account time lags. However, this solution is only valid for linear correlations [48] and it cannot be used with different correlation algorithms.

One of the constraints of measuring correlations is the need of a minimum sample
290 size to discover the correlations. If the number of observations is too low, the correlation algorithms could give false positives or false negatives. The sample size depends on the type of data and the data collection infrastructure. Therefore, for each type of data a preprocessing is needed to choose the accurate number of observations. It is also very helpful to have background knowledge such as understanding how the daily events can
295 affect the distribution of data. In big data analytics it is important to incorporate prior information that may be available about a domain in order to improve the results [49], [50]. For example, if we want to model the patterns of cars in a city, we know that the traffic follows a daily pattern, with more traffic at rush hours (beginning and ending of the general workday) and less traffic at nights. We also know that the traffic follows
300 a weekly pattern, with more traffic during weekdays and less during the weekends or

Figure 4: A visual representation of geographical coordinates on Google Map for a pair of road traffic sensors provided by city of Aarhus, Denmark. The blue line connects both sensors that are at the ends of the blue line.

holidays. We may also know that the traffic can differ from road to road and from location to location. For example, some roads could be busier during holidays and some during working days. All these a priori information are useful to reach accurate and fast results. We can check if we have a relevant number of observations in a week data by checking that the correlation factor does not change substantially when performing correlations with data for one week and with data for several weeks. In the case we have a relevant number of samples for a day or for a week, we could use one day or one week data to perform the correlation analysis between the traffic of two roads.

Our proposal for spatio-temporal correlation analysis is presented in Figure 5. Using only the three last blocks we can obtain a spatial correlation. The first two blocks add the temporal component. For clarity reasons we will first explain the last three blocks to obtain the spatial correlation. We will revisit the figure to include the temporal component with the first two blocks. We support the explanation with examples used in the experiments with real data.

The city of Aarhus has deployed 135 bluetooth lightweight sensors (see Figure 1) on street light poles that are low-power consuming and take the power from the street lighting. Therefore the deployment is easy as it does not imply traffic disruptions and the maintenance is low. The configuration of the sensors are performed remotely. The sensors measure in all weather conditions and have built-in directional antennas, which enables accurate positioning of detected devices.

Every time a mobile device with the bluetooth active is passing by a sensor the phone is anonymously registered and a record is updated with a timestamp and location information. All this records are encrypted and transmitted via broadband mobile networks to a central server every 5 minutes. The municipality of Aarhus with this data has created an open data platform ⁶.

The dataset used in this paper consists of traffic data from the platform. The data is organised in pairs of sensors providing information every 5 minutes regarding the geographical location of both sensors, time-stamp, and traffic intensity such as average

⁶<http://www.odaa.dk>

Figure 5: Block diagram of the methodology followed to calculate data streams movements in the cities. The window size Δw is variable, and we have tried window sizes of one hour, 12 hours, 1 day, 1 week, one month and 4 months. The time difference Δt is also variable, and we have tried steps of $\Delta t=5$ minutes, from -60 minutes to 60 minutes.

speed and vehicle count between them. Figure 4 shows the location of a pair of sensors, situated at the ends of the blue line, shown on a Google Map⁷. We will call this space between a pair of sensors a "sector".

Suppose stream A corresponds to the temporal series of the number of cars in a sector A, and stream B is the temporal series of the numbers of cars in sector B (see Figure 18). We want to measure if there is any correlation between the two streams.

We also want to measure the correlation with the streams over time. For example, we measure the correlation between stream A at t and stream B at time $t + \Delta t$. We measure the correlation with different Δt , in order to find the maximum correlation. The maximum correlation will provide the time that an event taking place in stream A (or in sector A) will show a response in stream B (or in sector B).

Assuming we have sufficient data in each stream to divide it into windows and each window includes sufficient data to measure correlations, we propose to use overlapping windows in order to improve the measurements (see Figure 5). In our example, we need to check if we have sufficient traffic in one week. The data contains 2016 observations per week, and we show that this number is enough, because the correlation does not change substantially within weekly observations or with several month observations (as we will explain in Section 5.2 and Figure 12). If we have traffic data for several weeks, we propose to improve the robustness of the correlation by making several measurement and calculating the average of the results. We split the data streams into windows. In Figure 5 the block *window splitter* is shown with a window size of $\Delta w = 1$ week. We then try to detect the correlation between the streams (in our example the stream are the numbers of cars) in two sectors of the city for each week of data. The correlation of each window is calculated (block *correlation* in Figure 5) with a correlation algorithm such as the two algorithms presented in section 3, Pearson correlation and mutual information. We calculate the average correlation of all the week-window blocks as the final correlation of sectors (block *average correlation* in Figure 5). To improve the robustness of the average correlations we use overlapping windows. Using overlapping windows, we can reuse the

⁷<https://maps.google.com/>

Figure 6: 4 weeks of traffic data divided in 50% overlapped windows.

observations for several experiments, virtually increasing the number of observations used in the correlations. Figure 6 shows the plot of 4 weeks data of the number of cars in one sector, this can be seen as stream A in Figure 5. We have divided this data to 50% overlapped windows of one week each. As we can see, window2 is 50% overlapped with window1 and 50% overlapped with window3.

With this procedure we capture the spatial correlation of the data streams and can infer how the different streams are spatially correlated. However, a temporal component is also needed to show how the patterns of a stream (or sector) are related to another stream over time. For this purpose, before performing the windowing and correlation analysis, we shift one data stream with respect to the other. In Figure 5 this is shown as blocks of *Time shifters*. *Stream B* is shifted a period of time (Δt) with respect to *stream A*. The functioning of block *Time shifter* is better explained in Figure 7 and Figure 8. For clarity reasons we have represented two simple streams in Figure 7. We keep stream A as it is and shift stream B over $\Delta t = -1$ and obtain Figure 8. The idea is to analyse any possible correlation between streams A and B in Figure 7, and then between streams A and B in Figure 8. We then calculate the maximum of the two correlations. We repeat the process with different Δt and at the end obtain the $\Delta t = \Delta t_m$ that gives the maximum correlation between these two streams. We can conclude that what happens in *stream A* at a time t will be reflected in *stream B* after a time Δt_m . In this way, we can have a clear idea of how the streams are correlated. In our example, the data is captured every 5 minutes, so we start with a $\Delta t = -60 \text{ minutes}$, and make an iteration every 5 minutes until it reaches to a $\Delta t = +60 \text{ minutes}$. We have chosen a start and end Δt that gives a sufficient time interval ($\pm 1 \text{ hour}$) to detect the correlations. The pseudo-code of the algorithm is shown in algorithm 1.

5. Use cases

To evaluate our proposal we have performed three sets of experiments. The first two sets of experiments are comparisons of Pearson correlation and mutual information. These two sets allow us to show the advantages and disadvantages of each algorithm and provide guidelines on which algorithm to use with respect to the target application

Figure 7: StreamA and StreamB to be correlated in time t .

Figure 8: StreamA in time t and StreamB in time $t-1$ to be correlated.

ALGORITHM 1: A pseudo code for discovering patterns of city flows

Input: Road sensory data divided in sectors (Sa, Sb, ...) and window size (Δw)

Output: The lag times giving max. correlations between two streams.

$\Delta t \in [-tmax, tmax]$

```

repeat
  for each pair of sectors  $Sx, Sy$  do
    for each  $\Delta t$  do
      for each  $\Delta w$  do
         $corr(i) = \text{correlation } Sx^{\Delta w}(t), Sy^{\Delta w}(t + \Delta t)$ 
      end
       $MeanInTime(j) = \text{mean}(corr)$ 
    end
     $\max(MeanInTime)$ 
  end
until  $EndOfData$ ;

```

Table 1: Simulated Signals

SIGNAL	DEPENDENCY
Y0	Random
Y1	Random
X0	Random normally distributed
X1	$a * X0 + b$
X2	$a * X0^2 + b$
X3	$a * X0^2 + bX0 + c$
X4	$a/X0 + b$
X5	$a * X0^2 + bX0 + c/X0 + d$

and available data. First we evaluate the performance of Pearson correlation and mutual information with synthetic data. We then perform the same evaluation with real data in the second set of experiments. Finally, in the third set of experiments, we apply the proposed methodology in algorithm 1 to traffic data in a real city environment. The next subsections explain the experiments in detail.

5.1. Datasets and Experiment set-ups

The synthetic data are a set of eight different streams of 1.000.000 samples each (see Table 1). These data is used for the first set of experiments.

The real data sets are collected from the public traffic data⁸ in the city of Aarhus. These data are explained in the previous section and are used for the second and third sets of experiments. For reproducibility in other cities, the sensors have to be deployed, preferably on the roadside light poles, to directly take the power from them, or they can be used as stand-alone with small solar panels.

⁸<http://www.odaa.dk/dataset/realtds-trafikdata>

Figure 9: Scatter plots of the synthetic traffic.

5.2. Evaluation

The first set of experiments consists of evaluation of the different performance of Pearson correlation and mutual information with synthetic traffic. Using synthetic data we can predefine and modify dependency of the streams. This evaluation will provide insight into when to use each of the algorithms and will show how the window size affect differently to both algorithms with different stream dependencies. The first experiment implies the correlation between 2 independent streams, $Y0$ and $Y2$, and the rest of experiments were performed with streams that follow some dependency between them ($X0$ with $X1, X2, X3, X4$ or $X5$). The dependency of the signals in each experiment could be easily seen in the scatter plots in Figure 9.

For this experiments we use the *window splitter* block, the *correlation* block and the *average correlation* block shown in Figure 5, and explained in section 4, but we do not use the blocks *time shifter*. For window sizes bigger than 10.000 samples, we can see the results in Figures 10 and 11. The maximum entropy of individual signals are the independent signals ($Y0, Y1, X0$) with a value of 4,364 and the rest of individual entropies are $H(X1) = 4,364, H(X2) = 3,1734, H(X3) = 3,1849, H(X4) = 0,3518, H(X5) = 0,3981$. The figures show that mutual information can detect quadratic dependencies like the correlation between $X0$ and $a * X0^2 + b$ or the correlation between $X0$ and $a * X0^2 + bX0 + c$, both of them with a value of 2,5, being maximum possible mutual information the $\min(H(X0), H(X2)) = 3,1734$ and $\min(H(X0), H(X3)) = 3,1849$. Mutual information results find more correlations compared to Pearson correlation. However, Pearson correlation does not detect quadratic correlations. We can observe as well, that none of them can detect division dependencies such as $X0$ and $a/X0 + b$ or $X0$ and $a * X0^2 + bX0 + c/X0 + d$, because both correlations are zero. Both algorithms, however, detect the independence of the random variables (correlation zero) and the linear dependency ($X0$ with $a * X0 + b$) with the maximum value of the correlation, in Pearson with a value of 1 and in mutual information with a value close to $\min(H(X0), H(X4))$ or $\min(H(X0), H(X5))$ in each case .

Figure 10: Mean of the results of Pearson correlation performed with Synthetic data.

Some facts that we could observe performing variations in the window size are that different streams perform differently. For streams with division dependency (X0 with X4 or X5), Pearson correlation does not detect any correlation (with mean zero and variance nearly zero) with any window size, from 1.000.000 to 1.000. But mutual information shows a great variance. It decreases with the window size increases. For example for windows of 100 samples the average correlation is 1 with high variance, for windows of 10.000 the average correlation is around 0.4 and for windows of 100.000 the average correlation is 0.25, only with bigger windows it stabilizes around zero with small variance.

For streams with quadratic dependence (X0 with X2 or X3) even that Pearson does not detect the correlation but mutual information does detect these correlations. Both correlations methods show a small variance, with all the window sizes. The variance does not specify any correlation but Pearson correlations have the higher variance with these kind of data streams. For mutual information these variances are much smaller than for division dependency streams.

Figure 11: Mean of the results of mutual Information performed with Synthetic data.

Pearson correlation is consistent with different window sizes of the same data streams. The mean of the correlation coefficient is almost the same for different window sizes. This is true for all except for quadratic relations. Mutual information also shows some variations for quadratic dependencies, and higher variations for division dependencies. Our explanation to this behaviour is based on the fact that Mutual Information requires larger data sets than the Pearson coefficient to be more precise [39], and therefore for small window sizes the variation of Mutual information is higher than that of Pearson Correlation.

The second set of experiments are performed to compare the performance of Pearson correlation and mutual information with real data. For this purpose, we used the traffic data described in Section 4. We perform an exploratory data analysis through a visual check [51] on the correlation of two variables "*average speed*" and "*vehicles count*", with the aid of scatter plots, such as the one shown in Figure 13.

During the experiments we tried different window sizes. With the Aarhus traffic data, a weekly window seems to be the optimum. To prove that the window sizes are optimal, we perform correlations with window sizes of 1 hour, 12 hours, 1 day, 1 week, 1 month and several months. We did not find any substantial differences in the results of the correlations between 1 week and higher windows. Figure 12 shows the results. These results do not try to compare Pearson correlation with mutual information performance. They provide a mean to determine the optimal window for which the correlation is stabilized in each case. For mutual information the mean of the correlation is higher for short window sizes and it decreases as the window size increases. The value for mutual information stabilises around the 1 week window size. For Pearson correlation the mean

of the correlation stabilises around 1 day, but still it has a high variation, whereas for one week window size the variation is also small. We try daily patterns as we predict that the traffic patterns follow a daily routine (i.e. background knowledge), with higher traffic at rush hours and lower traffic during the night. The weekly windows were based on the assumption that the weekly pattern could follow a lower traffic during weekends and higher traffic during weekdays. One day window seems to have not enough observations, but one week window is more accurate. When we have the window sizes defined, we split the data in windows, as shown in Figure 5, block *Window Splitter*. The correlations are computed with block *Correlation*, and finally the average correlation of all the windows, with block *Average Correlation* is calculated.

Figure 12: Influence of the window size in the correlation results. Boxplot of a sector measuring the correlation between Average speed and number of car, with different window sizes, with Pearson correlation coefficient and with mutual information. The figure shows that around 1 week window the correlations tend to stabilized.

The results obtained with Pearson correlation and with mutual information for window sizes of one week were similar in most of the cases. However, in some specific cases, mutual information had a higher accuracy. For example, in Figure 13 the scatter plot shows some degree of dependency, but Pearson correlation gives a correlation factor over the windows with mean close to zero and a small variance. For the same windows, mutual information gives a mean of 0.4464, which is a more valid value to be considered as the $\min(Entropy(speed), Entropy(numberofcars))$ is 2.0336 (see Figure 14). As expected, larger window sizes did not change the results, and smaller windows did not provide

Figure 13: Scatter plot between number of cars and average speed of the cars. With this data Pearson correlation fails to find the correlation.

enough data to perform the correlations, the results had a high variance and they did not show any correlations.

It is worth noting that we found a small percentage of cases where Pearson correlation do not detect at all the correlations and mutual information does. Other works in the field of traffic flow analysis show microscopic and macroscopic fundamental flow diagrams where the density and flow of vehicles always follows a characteristic curve [30]. However, this curve includes free flow, bound flow and congestion, and there can be streets that never reach the congestion or even the bound flow. Therefore Pearson may not detect the correlations depending on the linearity of the resulting curve in a data stream related to that particular street. Overall, in scatter plots with triangle that has a right angle such as the one shown in Figure 15, the Pearson measure tend to work better than in scatter plots with a triangle without right angle (Figure 13). In the last case, we can see that when the number of cars is higher, the velocity is not minimum. this means the street is not congested and the scatter plot is less linear and more similar to a quadratic signal. The latter is the main reason that Pearson measure fails to detect the correlation.

We are aware that mutual information has a higher time complexity than Pearson correlation. To verify and evaluate the impact of this, we performed performance evaluation using both algorithms. For the experiments we used a machine with 8 GB RAM memory and a 1.6 GHz processor in OSX version environment. The results shown in Figure 17 confirm this assertion. Figure 17 shows that mutual information in average

Figure 14: Boxplot of the results obtain by mutual information and Pearson correlation when evaluating the correlation between vehicle count and average speeds shown in Figure 13

is 3 times slower than Pearson correlation. The processing time for Pearson correlation is around 1 millisecond and mutual information is around 3 milliseconds. This can be improved by including part of the processing on the sensor nodes, or performing the correlation analysis off-line when real time analysis is not a strict requirement.

It is important to balance the computational effort with the accuracy of the results depending on the application restrictions and the datasets characteristics. For our datasets, if the accuracy of the application is important and it does not need to continuously update the model, a mutual information approach will be suitable. However, if an application prioritises real time response over accuracy, Pearson correlation will be suitable as it will only give a few false negatives. In other scenarios with different types of data streams (temperature, pollution, etc.) without a priori knowledge of the potential correlations it is better to use mutual information because we do not know the percentage of cases where Pearson correlation will fail to detect the correlations.

The third set of experiments evaluates our proposal to create a representation of changes and correlations in the data in a smart city environment, explained in section 4. We follow all the steps shown in Figure 5 and algorithm 1. The experiments calculate the time lag in observing one pattern of occurrence in one data stream and the impact of that occurrence on other data streams over time and space parameters. For example, an increase in traffic (due for example to a massive event) in one part of a street can show a correlation in the form of a traffic increase some time later in a different part of that road or in other related and/or linked roads. By using this elapsed time analysis we can find the spatio-temporal correlations. Most of the current correlation analysis solutions usually do one-to-one mapping and in real world data and events, one occurrence can be correlated to another occurrence in a different time/location. Our proposed method has

Figure 15: Scatter plot between number of cars and average speed of the cars. With this data Pearson correlation detects some correlation.

the advantage of analysing and detecting such correlations and dependencies.

We evaluate the feasibility of analysis of traffic data streams by making correlations elapsed in time. These experiments use only the *vehicle count* variable of the sectors. For example, an experiment is performed to identify the correlations in one avenue, between the variables *vehicle count* of two sectors of this street (see Figure 18). This correlations involved the decouple of both data streams (A and B) with different elapsed periods. We calculate the correlation of both streams $A(t)$ and $B(t)$. We shift one stream 5 minutes and calculate the correlation between $A(t+5 \text{ minutes})$ and $B(t)$. The 5 minutes time lag is chosen because the streams consists of readings of data every 5 minutes. We repeat the procedure for different lag window sizes from $t-60$ minutes to $t+60$ minutes. In this particular case, both correlation methods performs equally good, as $\min(Entropy(StreamA), Entropy(StreamB)) = 2,2573$ (see Table 2). The results show that the correlation is higher when we elapsed the data between these two sectors for 5 minutes. This indicates the traffic at one point of time in sector A will most probably have an impact on traffic in sector B after 5 minutes. Both correlation methods performed equally good, because in this case the scatter plot of both sectors follows some linearity (see Figure 19).

The experiments were extended to different sectors of the city, creating a view of the traffic data correlations. For example the data in sector A in Figure 20 and 19 has its maximum correlation with the data in sector B with a lapse of 5 minutes and it also has its maximum correlation with sector C with a lapse of 10 minutes. This implies that the traffic in sector A will have an impact on traffic in sector B after 5 minutes and in sector

Figure 16: Mean and variance of the results obtain by mutual information and Pearson correlation when evaluating the correlation between vehicle count and average speeds shown in Figure 15

Table 2: Table Measuring the Correlation between two sectors of one Avenue

TIME ELAPSED	AVERAGE M.I.	AVERAGE P.C.
-5 min	1.0899	0.4624
0 min	1.1317	0.4938
+5 min	1.1480	0.5195
+10 min	1.1099	0.4963
+15 min	1.0900	0.4740

C after 10 minutes. These analysis of the spatio-temporal correlations can help to show the dynamics of the city. For example, when we do not have an a priori knowledge of the city areas our analysis can provide an insight of the traffic flows, coupling together the roads that lead to working areas but decoupling them to the roads leading to leisure areas.

Our proposed method is based on creating and training a model with historical data. In terms of scalability for real time and large volumes of data, this model needs to be valid over time, and therefore the adaptation of the model with new historical data should be done only from time-to-time and in an off-line manner. The scalability problem could arise in monitoring the data in real-time in order to find outliers that differ from the model. The outlier detection techniques are out of the scope of this paper, but in general terms they need less processing time as the main idea is to calculate the difference between the actual value and the value in the model.

Figure 17: Comparison between the processing time of Pearson correlation and mutual information during the runs of each window.

6. Co-occurrence vs Causality: Towards an Analytical Examination of the Causation on Modelling City Movements

Big data analytics have been widely criticised due to the tendency to assume that every co-occurrence in the data underlies a causality. This criticism emerges from the use of correlations or co-occurrences to extract actionable information from big data. However, it is well-known that the relationship between co-occurrence and causality is not always bidirectional. When two or more variables are linked through a real causation it is always possible to find a relevant correlation factor between them. However, the existence of correlation between variables does not imply the existence of causality. There have been some assumptions on causality that have been proved wrong. For example, a study found a strong correlation between myopia and the use of night-lights in children, and the authors concluded that the exposure to night-lights causes myopia [52]. However, an a posteriori study showed that there is a strong correlation between children with myopia and parents with myopia, and that parents with myopia are more likely to use night-lights for their children. So, the study concluded that the myopia will probably have an inheritance causation rather than be caused by the exposure to night lights [53].

Several research has shown that human beings always try to find causations -for example extrapolating a word that kids hear just once and using it whenever they think is suitable- [54]. Scientific literature has criticized that human instinct to extrapolate or find causations of everything and tried to provide a more specific delimitation of causation and what makes it extremely difficult to find causations out of correlations. In particular it is assumed that casual relationships are greatly underdetermined by statistical relevance relationships [55]. Hence, several researchers try to avoid the word causation and even Karl Pearson proposed to replace the idea of causation entirely by the idea of correlation [56]. There are also several perspectives that prefer to refer to it as explanation instead of causation [57].

Figure 18: Location of two sectors of the city traffic in Aarhus in Denmark, sector A and B. In each sector there are two sensors, each situated in each icon. The data provided by each couple of sensors is the number of vehicles between both sensors, the average speed of the cars in the sector and other data like length of the sector, geolocation of the sensors, etc.

This paper is not trying to infer the causation of the correlation and therefore, it avoids the above mentioned criticism. Instead, our model provides a first step towards the delimitation of causation in the analysed relationships. One aim of the correlation analysis in smart cities is often to complete the missing data of faulty sensors with other information that have shown a correlation with the incomplete data recently. First, we use a case scenario and taken advantage of the recent correlation evidences in order to draw temporal views of data changes and correlations. Following this temporal correlations we try to decouple the causation by just following the recent correlations and not relying in long term correlations that can vary over the time.

Second, we use different procedures proposed in the statistical literature to make sure that the correlation may more likely to signal a causal relationship. Even though we do not try to find any causation, we believe that these guidelines help to develop good experiments that create reliable correlations. In this context, some works propose the use of randomised controlled experiments for discovering causal relationships. This technique is very popular in the field of medicine. In this method, the target population is selected and randomly split into two groups: one group will try the medicine and the other group will try a placebo. However, it is often unclear how to perform these experiments; which variables to use, and determine the number of experiments needed to identify the causation among the set of variables [58]. Some works try to define what will be good experiments (eg. [59]), but this is still an ongoing area.

Our solution has followed several of the most popular design guidelines in the literature to minimise the selection bias, the allocation bias, and the confounding effects of the correlation (see for example [60]). We have tried to delimitate a theory of change by

Figure 19: Scatter plot of the number of vehicles in both sectors A and B

610 noting the specific form and duration of change and predictors of change e.g. describing the repeated daily and weekly patterns. The utilisation of different time lags also helps to check the validity of the relationship, as we have tried different time windows. Finally, we have reported the best practices, rationals and conditions to use each correlation analysis method and have described their strengths and weaknesses.

615 Third, we propose the periodical re-calculation of the correlations between the different analysed flows of the city in order to minimise the effect of a misleading causation in the correlation. For example, we can calculate the flow of the traffic using weekly data.

Finally, if the correlation has a pattern, temporal series can be used to analyse if certain changes in the data are able to predict other values over a long period of time. 620 We also use longitudinal analysis that are much more precise for establishing causal relationships than cross-sectional designs (at one point in time) [61].

7. Discussion, Conclusions and Future Work

This paper has proposed a method to analyse spatio-temporal correlations in IoT and in particular in smart city road data streams. The novelty of this method lies in spatio- 625 temporal analysis and using the time elapsed between what is happening in a location and the impact that it has on another location. Previous works have dealt only with space or time separately. The work in the literature has mainly focused on how patterns are correlated at the same time or how the streams are related in a particular location. A secondary contribution of this paper has been the use of different statistical measures of dependability to analyse city data streams. We have compared Pearson correlation and 630 mutual information. We have described how the window size affect the dependency of different distributions of data and the balance between accuracy and computation cost

depending on the application restrictions and dataset characteristics. We have performed our experiments avoiding the misleading relationship between correlation and causation. Our analyses do not aim to infer causation, but to take advantage of recent correlations of data. We have also followed some guidelines in the literature to minimise the selection bias, the allocation bias, and the confounding effects of the correlation, not to infer causation but as good practises to developed accurate experiments.

Figure 20: Three sectors of the traffic data of Aarhus. The traffic data in sector A is correlated with the data of sector B with a lapse of 5 minutes, and it is also correlated with sector C with a lapse of 10 minutes.

Three sets of experiments have been performed, in the first one we have compared the performance of Pearson correlation and mutual information with synthetic traffic in which we have an a priori knowledge of any existing correlations (i.e. background knowledge). We have also used real data in which we did not have an a priori knowledge of any existing correlation. We have also experimented with real world traffic data to find the correlations between different windows of the data streams in different locations and over different time lapses. The data was taken from bluetooth lightweight sensors installed on light poles without disrupting the traffic. The sensors support all weather conditions, have low power consumption and need not maintenance as they can be configured remotely.

The study have demonstrated that for both, synthetic and real data, mutual information can discover dependencies in more general cases of distribution of data. Pearson correlation can discover linear distribution of data. However, mutual information requires more processing time. We have discovered that with real traffic mutual information in some cases performs better than Pearson correlations. It occurs in streets without traffic

congestion. Therefore, it is important to balance the accuracy with the computational
655 cost. We have also concluded that the sample size is important and it influences the
correlation methods. We have also used the proposed solutions to model the correlation
in smart city environments.

In summary, the paper first have compared two correlation algorithms, Pearson and
mutual information, studying window size selection criteria in streaming sensory data,
660 and evaluating the processing time and detection. Second it has proposed a model
to study the urban mobility with a spatio-temporal analysis of urban data using both
correlation techniques. And finally the experiments have followed some good practises
to decouple the correlations from the causations.

Our analysis contributes to optimize the use of the available infrastructures and to
665 better plan the future urban deployments. Urban developments planners should take
optimal decision on road expansion or addition based on minimizing travel times and
taking into account the effects of environmental sustainability [62], as well as the noise
[63]. To this aim a good knowledge of the movements inside the city is a must. In
particular, our analysis gives a view of the movements inside the city networks; how
670 different data related to people, traffic, electricity, etc. change around the city networks,
road network, or water and electricity in pipe networks. With a suitable view of these
changes city councils, planners, developers, industry and citizens can better manage and
effectively use the resources, make better planning of facilities and infrastructures, and
manage expected and unexpected events in the city. Some of these analysis, especially
675 for long term planning, will need a better study of the causality, and some of them will
rely on the correlation without questioning the causation. In particular, our use cases
can benefit city managers to better use road infrastructures, deriving correctly the traffic
when an event takes place in one part of the city, such a massive cultural event. Our
results can also be used to improve the infrastructures in the city by constructing or
680 extending parts of the road infrastructures.

Future work will focus on investigating the causation on the relationship, or correla-
tion, between two different events. If there is a correlation between event A and event B,
we want to investigate if A causes B or if B causes A, or find out if there is no causation.
Using spatial-temporal corrections and also considering the background knowledge, his-
685 torical data and contextual information, we will work to find more enhanced methods
and solutions to analyse the correlation versus causation. We also plan to annotate the
information in a semantic dynamic model [64] to better interoperate with other data
sources.

Acknowledgement

690 The authors would like to thank the constructive comments of the three reviewers
and editor. The research leading to these results has received funding from the European
Commission's in the H2020 for FIESTA-IoT project under grant agreement no. CNECT-
ICT-643943.

References

- 695 [1] United Nations, World urbanization prospects: The 2014 revision, highlights. department of eco-
nomic and social affairs, Population Division, United Nations. ISBN: 978-92-1-151517-6. [On-line;

- accessed 8-12-2017].
URL <https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Report.pdf>
- [2] Y. Zheng, Y. Liu, J. Yuan, X. Xie, Urban computing with taxicabs, in: Proceedings of the 13th international conference on Ubiquitous computing, ACM, 2011, pp. 89–98, ISBN: 978-1-4503-0630-0. doi:10.1145/2030112.2030126.
 - [3] T. Kindberg, M. Chalmers, E. Paulos, Guest editors' introduction: Urban computing, IEEE Pervasive Computing 6 (3) (2007) 18–20, ISSN: 1536-1268. doi:10.1109/MPRV.2007.57.
 - [4] I. Shklovski, M. F. Chang, Guest editors' introduction: Urban computing—navigating space and context, Computer 39 (9) (2006) 36–37, ISSN: 0018-9162. doi:10.1109/MC.2006.308.
 - [5] A. H. Alavi, A. H. Gandomi, Big data in civil engineering, Automation in Construction 79 (2017) 1–2, ISSN: 0926-5805. doi:<http://dx.doi.org/10.1016/j.autcon.2016.12.008>.
URL <http://www.sciencedirect.com/science/article/pii/S0926580516305246>
 - [6] R. G. Malgady, D. E. Krebs, Understanding correlation coefficients and regression, Physical therapy 66 (1) (1986) 110–120, ISSN: 0031-9023. doi:<https://doi.org/10.1093/ptj/66.1.110>.
 - [7] O. R. Ighodaro, P. Pascal, D. Virginia, Smart infrastructure: an emerging frontier for multidisciplinary research, Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction 170 (1) (2017) 8–16. doi:10.1680/jsmic.16.00002.
 - [8] D. Puiu, P. Barnaghi, R. Toenjes, D. Kümpfer, M. I. Ali, A. Mileo, J. X. Parreira, M. Fischer, S. Kolozali, N. Farajidavar, et al., Citypulse: Large scale data analytics framework for smart cities, IEEE Access 4 (2016) 1086–1108, ISSN: 2169-3536. doi:10.1109/ACCESS.2016.2541999.
 - [9] X. Zhan, S. V. Ukkusuri, C. Yang, A bayesian mixture model for short-term average link travel time estimation using large-scale limited information trip-based data, Automation in Construction 72 (2016) 237–246, ISSN: 0926-5805. doi:<https://doi.org/10.1016/j.autcon.2015.12.007>.
 - [10] P. Barnaghi, M. Bermudez-Edo, R. Tönjes, Challenges for quality of data in smart cities, Journal of Data and Information Quality (JDIQ) 6 (2) (2015) 6, ISSN: 1936-1955. doi:10.1145/2747881.
 - [11] N. Lathia, D. Quercia, J. Crowcroft, The hidden image of the city: sensing community well-being from urban mobility, in: Pervasive computing, Springer LNCS 7319, 2012, pp. 91–98, ISBN: 978-3-642-31204-5. doi:https://doi.org/10.1007/978-3-642-31205-2_6.
 - [12] D. Quercia, D. Ó. Séaghdha, J. Crowcroft, Talk of the city: Our tweets, our community happiness., in: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media -ICWSM, 2012, ISBN: 978-1-57735-556-4. [Online; accessed 8-12-2017].
URL <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4611/5056>
 - [13] D. Bollier, C. M. Firestone, The promise and peril of big data, Aspen Institute, Communications and Society Program Washington, DC, USA, 2010, ISBN: 0-89843-516-1, [On-line; accessed 8-12-2017].
URL <https://www.emc.co.tt/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf>
 - [14] J. H. Kell, I. J. Fullerton, M. K. Mills, Traffic detector handbook, Tech. rep., [On-line; accessed 8-12-2017]. (1990).
URL <https://www.fhwa.dot.gov/publications/research/safety/ip90002/ip90002.pdf>
 - [15] K. Kanistras, G. Martins, M. J. Rutherford, K. P. Valavanis, Survey of unmanned aerial vehicles (uavs) for traffic monitoring, in: Handbook of unmanned aerial vehicles, Springer, 2015, pp. 2643–2666, ISBN: 978-1-4799-0815-8. doi:10.1109/ICUAS.2013.6564694.
 - [16] L. Wang, F. Chen, H. Yin, Detecting and tracking vehicles in traffic by unmanned aerial vehicles, Automation in construction 72 (2016) 294–308, ISSN: 0926-5805. doi:<https://doi.org/10.1016/j.autcon.2016.05.008>.
 - [17] S. Kim, et al., Forecasting short-term air passenger demand using big data from search engine queries, Automation in Construction 70 (2016) 98–108, ISSN: 0926-5805. doi:<https://doi.org/10.1016/j.autcon.2016.06.009>.
 - [18] Y. Zheng, F. Liu, H.-P. Hsieh, U-air: When urban air quality inference meets big data, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 1436–1444, ISBN: 978-1-4503-2174-7. doi:10.1145/2487575.2488188.
 - [19] S. Kolozali, D. Puschmann, M. Bermudez-Edo, P. Barnaghi, On the effect of adaptive and nonadaptive analysis of time-series sensory data, IEEE Internet of Things Journal 3 (6) (2016) 1084–1098, ISSN: 2327-4662. doi:10.1109/JIOT.2016.2553080.
 - [20] J. Yuan, Y. Zheng, X. Xie, Discovering regions of different functions in a city using human mobility and POIs, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 186–194, ISBN: 978-1-4503-1462-6. doi:10.1145/2339530.2339561.

- [21] T. H. Savage, H. T. Vo, Yellow cabs as red corpuscles, in: IEEE International Conference on Big Data, 2013, IEEE, 2013, pp. 22–28, ISBN: 978-1-4799-1293-3. doi:10.1109/BigData.2013.6691773.
- [22] M. A. Yazici, C. Kamga, A. Singhal, A big data driven model for taxi drivers' airport pick-up decisions in new york city, in: IEEE International Conference on Big Data, 2013, IEEE, 2013, pp. 37–44, ISBN: 978-1-4799-1293-3. doi:10.1109/BigData.2013.6691775.
- [23] N. Ferreira, J. Poco, H. T. Vo, J. Freire, C. T. Silva, Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips, IEEE Transactions on Visualization and Computer Graphics 19 (12) (2013) 2149–2158, ISSN: 1077-2626. doi:10.1109/TVCG.2013.226.
- [24] H. Dong, M. Wu, X. Ding, L. Chu, L. Jia, Y. Qin, X. Zhou, Traffic zone division based on big data from mobile phone base stations, Transportation Research Part C: Emerging Technologies 58 (2015) 278–291, ISSN: 0968-090X. doi:https://doi.org/10.1016/j.trc.2015.06.007.
- [25] D. N. Bristow, C. A. Kennedy, Maximizing the use of energy in cities using an open systems network approach, Ecological Modelling 250 (2013) 155–164, ISSN: 0304-3800. doi:https://doi.org/10.1016/j.ecolmodel.2012.11.005.
- [26] H. Weisz, J. K. Steinberger, Reducing energy and material flows in cities, Current Opinion in Environmental Sustainability 2 (3) (2010) 185–192, ISSN: 1877-3435. doi:https://doi.org/10.1016/j.cosust.2010.05.010.
- [27] Y. Yang, K. K. Wong, Spatial distribution of tourist flows to china's cities, Tourism Geographies 15 (2) (2013) 338–363, ISSN: 1470-1340. doi:https://doi.org/10.1080/14616688.2012.675511.
- [28] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, S. H. Strogatz, Redrawing the map of great britain from a network of human interactions, PloS one 5 (12) (2010) e14248, ISSN: 1932-6203. doi:https://doi.org/10.1371/journal.pone.0014248.
- [29] N. Geroliminis, C. F. Daganzo, Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings, Transportation Research Part B: Methodological 42 (9) (2008) 759–770, ISSN: 0191-2615. doi:https://doi.org/10.1016/j.trb.2008.02.002.
- [30] C. F. Daganzo, N. Geroliminis, An analytical approximation for the macroscopic fundamental diagram of urban traffic, Transportation Research Part B: Methodological 42 (9) (2008) 771–781, ISSN: 0191-2615. doi:https://doi.org/10.1016/j.trb.2008.06.008.
- [31] J. P. Schorr, S. H. Hamdar, S. Kang, K. Jang, H. Kachouch, From structural equation modeling to macroscopic fundamental diagrams:: investigating the impact of road segments safety on network level efficiency, in: 17th International Conference Road Safety On Five Continents (RS5C 2016), Rio de Janeiro, Brazil, 17-19 May 2016, Statens väg-och transportforskningsinstitut, 2016, pp. 1–17, [On-line; accessed 8-12-2017].
URL <http://urn.kb.se/resolve?urn=urn:nbn:se:vti:diva-10619>
- [32] D. Ngoduy, R. Wilson, Multianticipative nonlocal macroscopic traffic model, Computer-Aided Civil and Infrastructure Engineering 29 (4) (2014) 248–263, ISSN: 1467-8667. doi:10.1111/mice.12035.
- [33] B. Heilmann, N.-E. El Faozi, O. de Mouzon, N. Hainitz, H. Koller, D. Bauer, C. Antoniou, Predicting motorway traffic performance by data fusion of local sensor data and electronic toll collection data, Computer-Aided Civil and Infrastructure Engineering 26 (6) (2011) 451–463, ISSN: 1467-8667. doi:10.1111/j.1467-8667.2010.00696.x.
- [34] L. Li, X. M. Chen, Vehicle headway modeling and its inferences in macroscopic/microscopic traffic flow theory: A survey, Transportation Research Part C: Emerging Technologies 76 (2017) 170–188, ISSN: 0968-090X. doi:https://doi.org/10.1016/j.trc.2017.01.007.
- [35] Y. Zheng, L. Capra, O. Wolfson, H. Yang, Urban computing: concepts, methodologies, and applications, ACM Transaction on Intelligent Systems and Technology (ACM TIST) ISSN: 2157-6904. doi:10.1145/2629592.
- [36] L. Li, X. Chen, L. Zhang, Multimodel ensemble for freeway traffic state estimations, IEEE Transactions on Intelligent Transportation Systems 15 (3) (2014) 1323–1336, ISSN: 1524-9050. doi:10.1109/TITS.2014.2299542.
- [37] M. Asif, J. Dauwels, C. Y. Goh, A. Oran, E. Fathi, M. Xu, M. Dhanya, N. Mitrovic, P. Jaillet, Spatiotemporal patterns in large-scale traffic speed prediction, IEEE Transactions on Intelligent Transportation Systems 15 (2) (2014) 794–804, ISSN: 1524-9050. doi:10.1109/TITS.2013.2290285.
- [38] T. Tchakian, B. Basu, M. O'Mahony, Real-time traffic flow forecasting using spectral analysis, IEEE Transactions on Intelligent Transportation Systems 13 (2) (2012) 519–526, ISSN: 1524-9050. doi:10.1109/TITS.2011.2174634.
- [39] J. Numata, O. Ebenhöf, E.-W. Knapp, Measuring correlations in metabolomic networks with mutual information, in: Genome Inform, Vol. 20, World Scientific, 2008, pp. 112–122, ISBN: 978-1-84816-299-0. doi:https://doi.org/10.1142/9781848163003_0010.
- [40] A. L. Edwards, An introduction to linear regression and correlation., WH Freeman, 1976, ISBN:

- 0716705613.
- [41] F. Kenney John, Mathematics Of Statistics, D. Van Nostrand Company, Inc; Toronto; New York; London, 1939, ISBN: 1258768720.
- [42] P. R. Bevington, D. K. Robinson, J. M. Blair, A. J. Mallinckrodt, S. McKay, et al., Data reduction and error analysis for the physical sciences, Vol. 7, AIP Publishing, 1993, ISBN: 0-07-911249.
- [43] J. Cohen, P. Cohen, S. G. West, L. S. Aiken, Applied multiple regression/correlation analysis for the behavioral sciences, Routledge, 2003, ISBN: 9780805822236.
- [44] B. Thompson, Canonical correlation analysis: Uses and interpretation, no. 47, Sage, 1984, ISBN: 0-8039-2392-9.
- [45] R. M. Gray, Entropy and information theory, Springer Science & Business Media, 2011, ISBN: 9781441979698.
- [46] R. M. Gray, Entropy and information theory, Vol. 1, Springer, 1990, ISBN: 978-1-4757-3984-8.
- [47] P. Barnaghi, A. Sheth, C. Henson, From data to actionable knowledge: Big data challenges in the web of things [guest editors' introduction], IEEE Intelligent Systems 28 (6) (2013) 6–11. doi: 10.1109/MIS.2013.142.
- [48] S. A. Billings, Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains, John Wiley & Sons, 2013, ISSN: 978-1-119-94359-4.
- [49] P. Niyogi, F. Girosi, T. Poggio, Incorporating prior information in machine learning by creating virtual examples, Proceedings of the IEEE 86 (11) (1998) 2196–2209, ISSN: 0018-9219. doi: 10.1109/5.726787.
- [50] M. Faulkner, R. Clayton, T. Heaton, K. M. Chandy, M. Kohler, J. Bunn, R. Guy, A. Liu, M. Olson, M. Cheng, et al., Community sense and response systems: Your phone as quake detector, ISSN: 0001-0782 (2013). doi:10.1145/2622633.
- [51] F. J. Anscombe, Graphs in statistical analysis, The American Statistician 27 (1) (1973) 17–21, ISSN: 0003-1305. doi:10.1080/00031305.1973.10478966.
- [52] G. E. Quinn, C. H. Shin, M. G. Maguire, R. A. Stone, Myopia and ambient lighting at night, Nature 399 (6732) (1999) 113–114, ISSN: 1476-4687. doi:doi:10.1038/20094.
- [53] K. Zadnik, L. A. Jones, B. C. Irvin, R. N. Kleinstein, R. E. Manny, J. A. Shin, D. O. Mutti, Vision: Myopia and ambient night-time lighting, Nature 404 (6774) (2000) 143–144, ISSN: 1476-4687. doi:doi:10.1038/35004661.
- [54] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, How to grow a mind: Statistics, structure, and abstraction, science 331 (6022) (2011) 1279–1285, ISSN: 0036-8075. doi:10.1126/science.1192788.
- [55] L. Sklar, Physical theory: method and interpretation, Oxford University Press, 2014, ISBN: 9780195145649 019514564X.
- [56] P. Spirtes, C. N. Glymour, R. Scheines, Causation, prediction, and search, Vol. 81, MIT press, 2000, ISBN: 0-262-19440-6.
- [57] H. Kincaid, The Oxford Handbook of Philosophy of Social Science, Oxford University Press, 2012, ISBN: 9780195392753.
- [58] A. Hyttinen, F. Eberhardt, P. O. Hoyer, Experiment selection for causal discovery, The Journal of Machine Learning Research 14 (1) (2013) 3041–3071, ISSN: 1532-4435. . [On-line; accessed 8-12-2017].
URL <http://resolver.caltech.edu/CaltechAUTHORS:20140123-114755984>
- [59] F. Eberhardt, Almost optimal intervention sets for causal discovery, in: UAI'08 Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, 2012, pp. 161–168, ISBN: 0-9749039-4-9.
- [60] R. E. Ployhart, R. J. Vandenberg, Longitudinal research: The theory, design, and analysis of change, Journal of Management 36 (1) (2010) 94–120, ISBN: 01492063. doi:<https://doi.org/10.1177/0149206309352110>.
- [61] T. W. Taris, M. A. Kompier, Cause and effect: Optimizing the designs of longitudinal studies in occupational health psychology, Work & Stress 28 (1) (2014) 1–8. doi:<https://doi.org/10.1080/02678373.2014.878494>.
- [62] Y. Wang, W. Y. Szeto, Multiobjective environmentally sustainable road network design using pareto optimization, Computer-Aided Civil and Infrastructure Engineering 32 (11) 964–987, ISSN: 1467-8667. doi:10.1111/mice.12305.
URL <http://dx.doi.org/10.1111/mice.12305>
- [63] S. Abo-Qudais, A. Alhiary, Statistical models for traffic noise at signalized intersections, Building and Environment 42 (8) (2007) 2939–2948. doi:<https://doi.org/10.1016/j.buildenv.2005.05.040>.

- 875 [64] M. Bermudez-Edo, T. Elsaleh, P. Barnaghi, K. Taylor, Iot-lite: a lightweight semantic model for the internet of things and its use with dynamic semantics, *Personal and Ubiquitous Computing* (2017) 1–13doi:10.1007/s00779-017-1010-8.
URL <http://dx.doi.org/10.1007/s00779-017-1010-8>