Automation in Construction 88 (2018) 87–100

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Analysing real world data streams with spatio-temporal correlations: Entropy vs. Pearson correlation

Maria Bermudez-Edo[a,*], Payam Barnaghi[b], Klaus Moessner[b]

[a] University of Granada, Granada, Spain
[b] University of Surrey, Guildford, UK

ARTICLE INFO

Keywords:
Smart cities
Internet of things
Correlation
Entropy

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.

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 [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 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 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 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 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 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 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 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 optimize the operational efficiency, and this area will need a combination of efforts from different research fields [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

* Corresponding author.
E-mail address: mbe@ugr.es (M. Bermudez-Edo).

https://doi.org/10.1016/j.autcon.2017.12.036
Received 15 May 2017; Received in revised form 12 December 2017; Accepted 29 December 2017

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 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 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 statistics some alternatives are available such as distance correlation, mutual information or correlation ratio. Although none of the latter solutions are completely accurate, these 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 dependencies. We analyse both approaches, first with synthetic data and then with real data taken from the smart open data Aarhus platform[1] 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 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 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 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 Fig. 1). These sensors are easier to deploy than traditional transportation sensors located on the ground, because they can 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 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 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 causation.

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

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.

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 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 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 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 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 consuming [16].

Taking advantage of the sensor deployments in smart cities several recent studies applied big data to urban infrastructure operations [5,17-

[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.