An Assessment of Sensor Locations Based on Data Analysis Tools

Article · June 2016

CITATIONS
0

READS
36

2 authors:

Khalid Kahloot
Budapest University of Technology and Economics
11 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Péter Ekler
Budapest University of Technology and Economics
80 PUBLICATIONS **167** CITATIONS
SEE PROFILE

An Assessment of Sensor Locations Based on Data Analysis Tools

Khalid Kahloot Peter Ekler
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
{Kahloot.Khalid,Ekler.Peter}@aut.bme.hu

Abstract. In the urban environments, it is important to assure that the resources are exploited in an optimal way. In smart cities, a network of outdoor sensors connected with remote control systems are used to monitor, control, and optimize operating infrastructure such as transportation systems. With respect to sensor deployment, analysts keep working on collected data to aid the optimal relocations of the sensors. Correct deploying of sensors is an important factor that helps reading actual data in real time without redundancy or noise. System operators should maintain frequent redeployment plan for sensors in order to optimize the control systems. Such plans are much convincing if they rely on data analysis of historical records of sensors records.

In this paper, a data analysis case study is presented related to the SOSLUN project. SOLSUN is an urban network, deployed in Loddon, UK. In this network, sensors are attached to public lightings that measure air quality, noise level and other values. This research carried out different data analysis methods based on these sensor values. Different association, clustering and dimension reduction algorithms were applied and the results are summarized in this paper.

Keywords: Data Analytics, SOLSUN, Smart City, Street Lighting Control, Internet of Things

1 Introduction

The SOLSUN (Sustainable Outdoor Lighting and Sensory Urban Network) project use street lighting control systems to provide data and connectivity for smart city applications. The solution stands from different sensor modules that are installed on the public lightings around the city. These modules contain several sensors like electric sensors, pollution sensors, and light sensors. Loddon, in the UK, is the first town where the SOLSUN technology was deployed. The group of sensor modules create a self-healing mesh network that basically connects the light posts together. It is planned that the solution will be deployed in Budapest as well in the upcoming years.

The main concept behind the solution is that sensors installed on the street lights can collect environmental data and also to reduce energy consumption by controlling the lights in a smarter way. For example if there is no traffic in the area, the light level can be decreased. This solution also provides a new type of network from street lights and other applications can be created on

this platform. Sensors on the street lights capture data like air pollution, noise pollution, traffic density, etc. The information that is gathered by the system can be used to address traffic congestion or monitor GHG (Green House Gases) emissions in the city. The data, collected by the sensors are used basically to control different type of smart systems around the city. Furthermore, this data can be also used to study and understand different patterns in the city.

The surface and atmospheric modifications due to urbanization generally lead to some changes that can be detected via these sensors. For example, it was shown earlier that urbanized areas are generally warmer than the surrounding non-urbanized areas [1]. This change in light, temperature and electricity consumption can be observed via SOLSUN. These negative changes can be reduced by a smart lighting system when for example the burn hours and energy consumption is optimized based on the traffic as mentioned earlier.

In this paper, we have monitored the deployed SOLSUN system in Loddon and we introduce measurements and data analytics on the collected data. The rest of this paper is organized as follows. Section 2 introduces previous related work in the field of smart cities and environmental measurements. Section 3 describes the available data set and open questions. Section 4 gives an overview about the measurement methodology that included sampling, normalization and association. Section 5 discusses about the used methods, clustering and the results via figures. Section 6 concludes the paper and proposes further research directions.

2 Related Work

This section introduces the work of other researchers related to environmental data analysis. In [1], Voogt and Oke were considering thermal data. They examined three types of researches [2], [3], [4]. They presented the spatial structure of urban thermal patterns and their relation to urban surface characteristics. In [4], Bendor and Saaroni studied the relation between atmospheric heat islands and surface urban heat islands. They studied thermal remote sensing over vegetated surfaces to the study of urban areas. In [2], [5] and [6], they draw a strong association between temperature and normalized difference vegetation index. The negative relationship is that higher levels of latent heat fluxes are more representative in areas like forest and grassland. Wilson and Langley in [6] used Los Angeles and Paris metropolises as study cases to analyze physical processes, which determine energy fluxes and their interaction at the urban surface with multi-sensor satellite data.

Zeller et. al. [7] used land cover types from the California Wildlife Habitat Relationship database as independent variables. There were 25 mapped land cover types present in the study area, but many types had very low occurrence. They recommended examining a continuum of scales and behavioral states when using point selection functions to estimate resistance.

Analysis of geographic variations in incidence or mortality in relation to exposure to environmental variables is a subject of great interest in public health.

Each variable involved requires special attention in order to be able to interpret associations meaningfully in the context of a regression model [8], [9], [10].

Measuring efficiency of operations in this type of environments is also an important subject. Based on collected data, operators can suggest optimization plans. Kimes [11] recommends the basic concept of perishable asset revenue management, which determines the optimal trade-off between average daily rates and occupancy rates. However, Wassenaar and Stafford [12] advocate the use of a lodging index indicator for the hotel/motel industry. The lodging index is defined as the average revenue realized from each room, vacant or occupied, within a region or city during a given time period.

Other researchers exceed analysis to prediction. E. Petre [13] was working on forecasting weather by using CART (Classification and Regression Trees) based on data collected in Hong Kong area. In [14], Cortez and Morais present an approach for predicting the burned area of forest fires by using meteorological data. They used recent real dataset, collected from the northeast region of Portugal. They apply several experiments with five data mining techniques (i.e. multiple regression, Decision Trees (DT), Random Forest (RF), Neural Networks (NN) and Support Vector Machine (SVM)).

Our research focuses on data collected from the SOLSUN system in Loddon city and analysis this data from different perspective to understand smart city concepts in more details.

3 Problem Statement

This section describes general challenges that should be considered related to urban data sets and introduces the available data set.

3.1 Challenges

One of the challenges that we have to consider is to choose the appropriate statistical tool for urban data. Some data analysis tools require to the special format of data while others are not efficient in some situations. Therefore first we need to do an exploratory analysis to determine factors such as means, maxima, and quantiles. There are common methods that can be used to discover association rules and calculate k-means, and t-SNE.

Before any complex data analysis can be performed it is important to execute the necessary transformations and standardize the values. There are both statistical and environmental reasons for considering adjustment of the data. We aim to reduce the effect of total quantity in sample units, to put the focus on relative quantities.

3.2 Description of Dataset

During this research we have carried out analysis over the data set, which was retrieved from SOLSUN database. It contains details about sites, containers,

devices, readings, and some calculations. The SOLSUN system provides collected data from sensors and it displays them in a standard format. The data in the system is generally useful for citizens and maintainers of the public lightings. This data can also be used to analyze and discover hidden correlations.

The available data set identify areas where sensors are installed. Currently there is only one site in the system for Loddon area, but the data structure supports multiple sites. The latitude and longitude are specified with a status flag, which indicates whether this site is active, warning, off-line, or unknown status. Inside each site, many containers are deployed. A container is installed on a lamp post and contains several sensors. The vendor of the sensors, the exact name and location information are displayed. Devices are stored in containers that can measure a set of parameters related to its environment. A device in the database is an actual sensor in the system.

Reading values in the database contain also time stamp labels. These readings can be auxiliary voltage, lamp current, lamp voltage, mains voltage, power, power supply unit current, temperature measured inside the container and light level. Besides, the database also contains some basic calculations like the number of hours when the lamps were switched on and the energy consumed by the lamps at the given day and also the green house gas emission around the container.

4 Methodology

This section describes the applied methodologies for our measurements and data analysis.

4.1 Exploratory Analysis

The initial step is to explore the data set by applying aggregations because the dataset is too big to look at as a whole. The dimensions of the available data set are 15 features with 1048575 records. The goal of this step is to identify problems and anomalies.

When summarizing the dataset, we found that more than 54 percent of the data measured by light and temperature sensors and the remaining reading came from electric sensors.

Table 1 shows the ranges of the readings along with the mean and median values. For temperature, the first quartile is 14.5 celsius and the third quartile is 31.5 celsius. For the light values, the first quartile is 0.2lx and the third quartile is 1927.5lx. To divide the range of the readings into contiguous intervals and to see the observations in a sample, in the same way, we applied the quantile function. We can see that 75 percent of values are above 14.9V in auxiliary voltage and above 14.9A in lamp current and above 27V in lamp voltage and above 45w in power and above 0.111V in the power supply current and above 403.7V in the power supply voltage. On the other hand, the values of main voltage, temperature and light are almost fairly distributed between quantiles.

For the mains voltage, 25 percent are less than 228.7V and 75 percent are above 232.5V. For the temperature, 25 percent are less than 14.5 celsius and 75 percent are above 31.5 celsius. For the light, 25 percent are less than 0.206lx and 75 percent are above 1927.5lx. Meanwhile, we have three calculated values from the collected readings. The first value is the energy consumed by each container, which has a mean of 338.23j. The second is the burning hours of lamps, which has a mean of 13.32 hours a day. The last is the greenhouse gases, which has a mean less than one.

Table 1: Readings ranges

Reading	**Max**	**Mean**	**Median**	σ	**n**
Auxiliary Voltage	15.5	6.7	0	7.41	3
Lamp Current	65.5	9.6	0	18	19
Lamp Voltage	177	12.4	0	44.25	117
Power	73.9	20.9	0	24	35
PSU Current	439	182.8	0	109.75	223
Temperature	1000	24.3	20.5	15	14
Light	1927.5	886.2	72.2	481.87	139
Energy	3157.19	338.23	133.31	412.51	47
Burn Hours	70	13.32	7.50	12	18
Burn Hours	70	13.32	7.50	12	18
GHG	1.459	0.156	0.062	0.1906	1

4.2 Sampling

Sampling can be particularly useful with data sets that are large like in this situation. First a a margin of error is selected and a the sample set is used to finalize the proper margin of error. The boundary of the confidence interval is expressed in equation 1.

The margin of error is the term right to Âą sign. Where the x is the sample mean and it is different from the value for . The z is used as the critical value. This value only depends on the confidence level of the test. The n is the size of sample data that we are estimating. In order to choose the appropriate size for the sample data, we use a hypothesis based on experiences. By choosing the margin of error e and choosing the confidence interval, most often 1.96 for 95 percent of confidence, we can solve the equation for n. After rearranging equation 1 and some basic changes, the sample size can be calculated as shown in equation 2.

$$x = \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{1}$$

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2} \tag{2}$$

In order to calculate the final value of the sample size, the standard deviation σ of the population is required. One option to estimate σ is based on previous studies or industry standard, that relied on the similar population of interest. A second option is to conduct a pilot study, e.g. survey, in order to select a preliminary sample. In statistical analysis, if the value of a parameter is not known, an empirically valid value can be used to substitute. Another option for choosing σ is to use judgment for σ, which is the data range divided by 4. Fortunately, there is *sd* function in *R* language for calculating the standard deviation.

In the dataset of SOLSUN, we calculated the estimated standard deviation for each reading as shown in Table 1. The confident interval was set to 95 percent with a margin error of 8. As mentioned in [13], each confident interval has a corresponding probability, which is calculated and included in the distribution table. In our case Z-score is 1 and α is *1 - confident interval* that is 0.05. To interpret the n column in Table 1, we will take the power as an example. To have 95 percent confident interval of the sample means to contain the population means μ each sample needs to be 35 readings. We should take a sample of 35 at the first time. Then we should take another sample and do that over and over again until we achieve a convergence.

4.3 Normalization and Outliers Detection

Normalization means adjusting values measured on different scales to a common scale. Following a matrix will be filled with random uniform variates between 0 and 1 and center and scale them to have 0 mean and unit standard deviation for each feature.

From another perspective, an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. We applied the outlier function over the normalized data set and we noticed that there are no such far away from values.

4.4 Association

Before applying association, first attributes were converted from numerical to binary. This step converts the attributes from float number to binomial value based on the third quantile because we want to highlight sensors readings. This means that all values above the third quartile will be represented by *TRUE* otherwise are they are *FALSE*. Association is the first step of data analysis. It aims to discover the association rules showing attribute-value conditions, which occur frequently together in a given set of data. Hidden relationships can be revealed between for example temperature and burning hours. In this paper, we used the apriori operator to generate association rules.

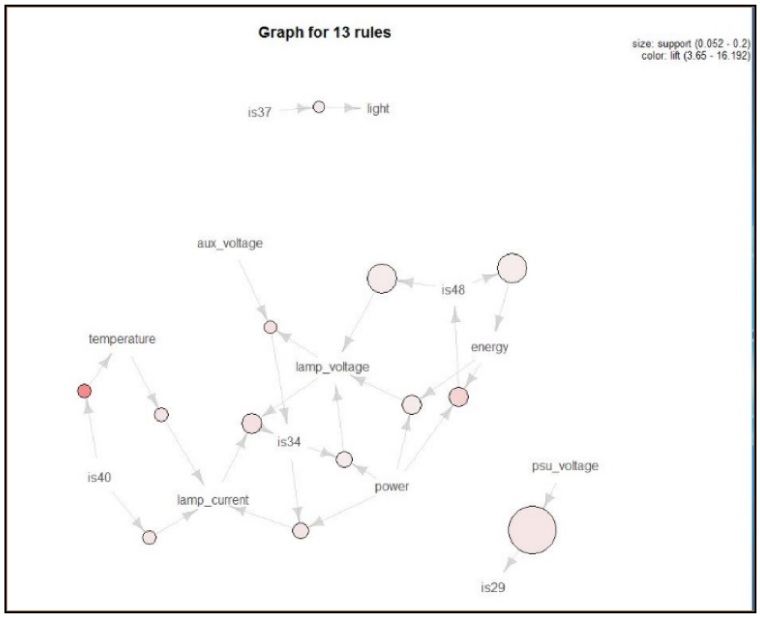

Figure 1: Association Visualization

5 Discussion and Analysis

This section contains the analysis of the data and discussion about the structure and value of it.

5.1 Association Results

The results of applying association rules showed 32 associations between containers and measured values. The container located in Beccles road measures the highest temperature and consumes the highest lamp current. Regardless of which location, the analysis showed that there is a strong correlation between temperature and lamp current. In other words, the higher temperature scores, the much lamp current consumed. The Beccles road location also scores high reading in PSU voltage and energy in another container. However, it also has strong the correlation between temperature and lamp current value.

The Gunton road location has two observations. In one container, there is a strong correlation between lamp voltage and auxiliary voltage from one side, and a correlation between lamp voltage and lamp current with respect to power readings. In fact, this correlation represents a logical relation between the three quantities since the can be derived from each other, however, the SOLSUN provides separate sensors for each one. Furthermore, in another container on the Gunton road location, the light scores high readings as well.

In the Low Bungay road location, one container has correlations of the readings. The analysis showed that lamp voltage and power and correlated with energy consumption. In addition, there is a correlation between lamp voltage and light with respect to energy consumption. Furthermore, association rules as visualized in Figure 1.

5.2 Clustering Results

After applying K-means clustering, five clusters can be observed as shown in Figure 2. Within a cluster, the sum of squares by cluster has a maximum value of 835.96 links two locations of Warren view and Beccles road. In the other hand, the minimum value of the sum of squares is 68.7 for the location of Cannell road. The centers of Clusters are locations of Gunton road, Beccles road, Freeman close, Cannell road and Alfric close respectively.

In this paper, we utilize cluster analysis, the process of grouping objects together in such a way that objects in one group are more similar than readings in other groups. For example, from data set of SOLSUN identifying containers with similar readings activities and group them together in clusters. Later these identified clusters can be targeted for business improvement by issuing special offers, etc.

In Figure 2, the clustering algorithm has grouped the input data into five groups. There are 3 Popular clustering algorithms, hierarchical cluster analysis, K-means cluster analysis and two-step cluster analysis. In this paper we deal with K-means clustering.

The first step is cluster assignment. Randomly, chose two cluster points (red dot & green dot) and assign each data point to one of the two cluster points whichever is closer to it. The second step is move centroid. In this step, we take the average of the points of all the examples in each group and move the Centroid to the new position i.e. mean position calculated. The above steps are repeated until all the data points are grouped into 5 groups and the mean of the data points at the end of the second step does not change.

In 2, we can see that the k-means algorithm cluster data into five clusters. The first cluster links two locations of Warren view and Beccles road. In this cluster, there is an association between temperature, lamp voltage and light. The second cluster located in Cannell road and the obvious readings are PSU voltage and lamp current. The third cluster located in Freeman close and it draws an association between auxiliary voltage and energy. The fourth cluster located in Beccles road with one observed reading that is auxiliary voltage. The fifth cluster located in Cannell road and it draws an association between lamp voltage and PUS voltage.

The size of the clusters organized as cluster one is the smallest and cluster three is the largest. Cluster two and cluster three has close sizes, but as shown in the figure, the cluster two has more variations of readings than cluster three. We can see that the location of Cannell road appears in two clusters. Moreover, the lamp current and voltage are the most focusing readings.

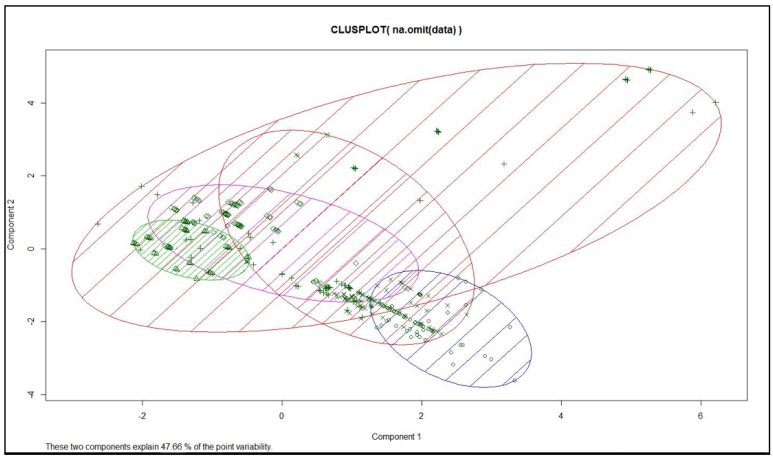

Figure 2: K-means Clustering

5.3 PCA and K-means Clustering

As this a data set of many variables, it will be an interesting opportunity to practice using a dimensionality reduction method to make the information easier to visualize and analyze. These variables are about in the middle of the data frame, so we can visualize all of them at once as shown is Figure 3.

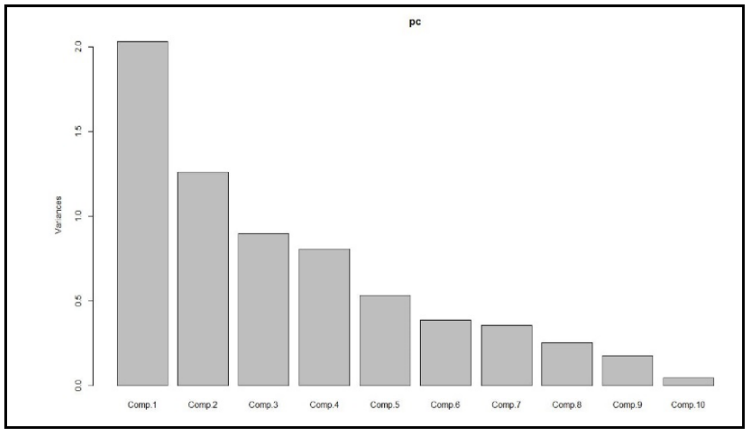

Figure 3: PCA Visualization

5.4 Dimension reduction with t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) [15] is a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. t-SNE or PCA are not a clustering method and they are being not necessarily a method primarily developed for dimension reduction. As shown in Figure 4, we can see features are overlaid. However, the most distinguished color is blue, which stands for temperature. In the addition, the highest container is 40, which located in Beccles road.

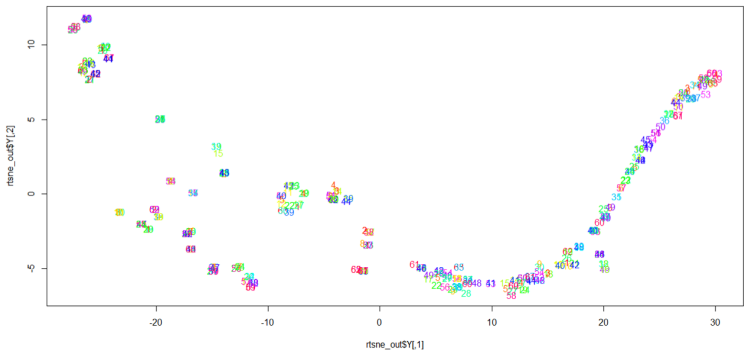

Figure 4: t-SNE Visualization

6 Conclusion

In this paper, we carried out data analysis tools over a dataset of urban environment. Data is measured by sensors, which are embedded inside containers in specific locations. Based on the available data from Loddon city, it can be concluded that Beccles road is the most active location. In addition, temperature and lamp voltage are the most scored values. In the other hand, the location of Alfric close is the most inactive location and GHG and auxiliary voltage is not that significant with respect with other readings. We recommend to relocate containers based on the analysis we provided. For example, the active locations should have more containers.

Future work includes retrieving more data from Loddon and other areas where SOLSUN will be deployed and develop prediction algorithms based on the available sensor data.

Acknowledgments

This work was supported by the Janos Bolyai Research Fellowship of the Hungarian Academy of Sciences.

References

[1] J. Voogt and T. Oke, "Thermal remote sensing of urban climates," in *Remote Sensing of Environment*, pp. 370–384, 2003.

[2] B. Dousset and F. Gourmelon, "Satellite multi-sensor data analysis of urban surface temperatures and landcover," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, pp. 43–54, June 2003.

[3] R. C. Balling and S. W. Brazel, "High-resolution surface temperature patterns in a complex urban terrain," *Photogrammetric Engineering and Remote Sensing*, vol. 54, pp. 1289–1293, 1988.

[4] H. Saaroni and E. Bendor, "Airborne video thermal radiometry as a tool for monitoring microscale structures of the urban heat island," *International Journal of Remote Sensing*, vol. 18, pp. 3039–3053, September 1997.

[5] K. P. Gallo and J. D. Tarpley, "The comparison of vegetation index and surface temperature composites for urban heat-island analysis," *International Journal of Remote Sensors*, vol. 17, pp. 3071–3076, June 1996.

[6] E. Wilson and E. Langley, "Methods for detecting domain interactions in nuclear receptors," *Research Support,U.S. Gov't, Non-P.H.S.*, vol. 364, pp. 142–152, 2003.

[7] K. Zeller, K. McGarigal, P. Beier, S. Cushman, T. Vickers, and W. Boyce, "Sensitivity of landscape resistance estimates based on point selection functions to scale and behavioral state: pumas as a case study," *Scientific Journal (JRNL), Landscape Ecology*, vol. 29, pp. 541–557, Maj 2014.

[8] D. G. Cook and S. J. Pocock, "Multiple regression in geographical mortality studies, with allowance for spatially correlated errors," *Biometrics*, vol. 39, pp. 361–371, 1983.

[9] S. Richardson, "Statistical methods for geographical correlation studies," in *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, pp. 181–204, 1992.

[10] S. Richardson and C. Monfort, "Ecological correlation studies," in *Spatial Epidemiology: Methods and Applications*, 2000.

[11] E. Kimes, "The basics of yield management," *Cornell Hotel and Restaurant Administration Quarterly*, vol. 30, pp. 14–19, 1989.

[12] J. Wassenaar and R. Stafford, "The lodging index: An economic indicator for the hotel/motel industry," *Journal of Travel Research*, vol. 30, pp. 18–21, 1991.

[13] "Checking out statistical confidence interval critical values - for dummies." `www.dummies.com`. Accessed: 2016-02-11.

[14] P. Cortez and A. Morais, "A data mining approach to predict forest fires using meteorological data," in *Proceedings of the 13th EPIA Â Portuguese Conference on Artificial Intelligence*, 2007.

[15] L. van der Maaten, "Accelerating t-sne using tree-based algorithms," *Journal of Machine Learning Research*, pp. 3221–3245, Oct 2014.