An Interactive Visualization System to Analyze and Predict Urban Construction Dynamics

Tzu-Chi Yen
Sensoro Technology Co., Ltd.,
Beijing, China
junipertcy@gmail.com

Tzu-Yun Lin
Academia Sinica,
Taipei, Taiwan
power6188@gmail.com

Ching-Yuan Yeh
National Taiwan University Hospital,
Taipei, Taiwan
jimteayeh@gmail.com

Hsun-Ping Hsieh
National Taiwan University
Taipei, Taiwan
sandoh714@gmail.com

Cheng-Te Li
Academia Sinica,
Taipei, Taiwan
ctli@citi.sinica.edu.tw

ABSTRACT

In this work, we develop an analytic and visualization system for citizens and government agencies to understand, track, and predict the construction dynamics in an urban area. The presented system consists of three major functions. First, we provide an interactive data visualization interface that allows various users to easily investigate the relationships among different construction types, geographical regions of interest, and the duration of construction over time. Second, the system enables the users to discover functional areas and identify possible urban construction problems by automatically grouping regions (i.e., villages) that share similar characteristics of construction requests. Third, a regression-based prediction model is devised to forecast the future numbers of construction requests for a certain construction type of interest and a specific region of interest. The one-year construction request data collected from Taipei City open data platform is used in our system. Our system is general-purposed for other data with similar schema, with a case demonstration in the urban construction data in this work. Researchers will be allowed to use our system to analyze other request data such as New York City 311 calls.

Categories and Subject Descriptors

H.2.8 [Information Systems]: Database Applications – *Data Mining, Spatial databases and GIS.* H.5.2 [Information Systems]: Information Interfaces and Presentation – *User Interfaces.*

General Terms

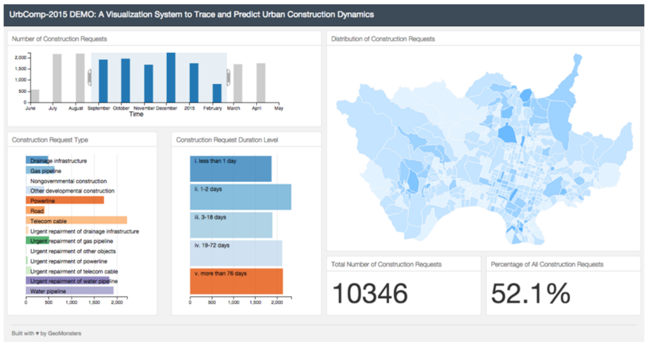
Measurement, Performance, Design, Human Factors.

Keywords

Construction requests, interactive visualization, urban computing, open data, regional analysis, clustering, regression

1. INTRODUCTION

With the maturity of information and communication technology, the government agencies of modern cities choose to build various online request platforms that are developed for either citizens or governors. The New York City 311 calling service[1] is created for citizens to reflect any problem they encountered in the urban environment. The Taipei Road Dig and Construction platform[2] is established for governors (e.g. different government departments and the companies of electronics/gas/water) to enable a more effective management of construction requests in urban areas. The requests are usually diversified with different types, and contain rich information of citizen/governor needs. Being a subcategory of urban requests, the road construction requests in Taipei are collected by the Public Works Department[3] and made public by the Taipei City Government. When a road section is subjected to closure due to certain reason, it will be applied and logged to the database after governmental approval. These logs reflect the needs of citizens at detailed space and time, including urgent repair (e.g. sudden water-pipe bursts) and scheduled construction (e.g. building a draining infrastructure). That says, the city construction request records can collectively depict urban construction events through not only their geographical and temporal dynamics of construction requests but also the corresponding labeled causes of construction events. These data can be analyzed by dynamically grouping regions that have the similar patterns of construction requests. For example, we might be able to identify some regions that both the powerline *and* water pipeline related constructions play the dominant cause of construction requests.



UrbComp-2015 DEMO: A Visualization System to Trace and Predict Urban Construction Dynamics

Total Number of Construction Requests
10346

Percentage of All Construction Requests
52.1%

Number of Construction Requests
2,000
1,500
1,000
500
0
June
July
August
September
October
November
December
2015
February
March
April
May
Time
Distribution of Construction Requests
Construction Request Type
Drainage infrastructure
Gas pipeline
Nongovernmental construction
Other developmental construction
Powerline
Road
Telecom cable
Urgent repairment of drainage infrastructure
Urgent repairment of gas pipeline
Urgent repairment of other objects
Urgent repairment of powerline
Urgent repairment of telecom cable
Urgent repairment of water pipeline
Water pipeline
0
500
1,000
1,500
2,000
Construction Request Duration Level
i. less than 1 day
ii. 1-2 days
iii. 3-18 days
iv. 19-72 days
v. more than 76 days
0
500
1,000
1,500
2,000
Built with ♥ by GeoMonsters

Figure 1: A snapshot of our system.

The analysis might also help to unveil the underlying problems of infrastructures or services. For example, the relatively high frequency of reporting in water-pipe burst in a certain region might hint aging or structural weakness of the water pipeline in that area. The prolonged durations of construction might indicate inefficient constructions. In addition to the visual presentation and analytics, the urban construction request data also offer some potentials to be used to build a predictive model that estimates the future construction requests of certain types in any urban area of interest.

UrbComp'15, August 10, 2015, Sydney, Australia.

[1] http://www.nyc.gov/311

[2] http://www.dig.tcg.gov.tw/tpdig/

[3] http://pwd.gov.taipei/

In this work, we aim to develop a novel system for presenting and analyzing the construction request data in Taipei City. Our system consists of three components. The first is *Interactive Data Processing and Visualization*, which not only prune noise data of construction requests but also provide an interactive manner for users to investigate the construction request data from various spatial and temporal granularity levels and different purposes of construction requests. The second is *Clustering-based Region Exploration*. We apply a clustering method to group geographical regions that share similar features of construction dynamics in the construction request data. From the results of clustering, users are expected to have insights about the city functions, the potential unsolved problems, and the distribution of requests in terms of construction dynamics. The third is *Construction Request Forecasting*. Based on the historical numbers of construction request data and some regional properties (e.g. the numbers of stations of Metro and bus), we aim at forecasting the number of construction requests for any region of interest so that the government agency is able to have a more effective and efficient resource management and deployment of urban construction.

We believe the developed system can provide three-fold benefits for various user sets. First, for citizens, the system can allow them to not only understand the true needs or the construction problems of their regional community, but also track and supervise the government agency in charge of the construction progresses. Second, for the city governors or public service providers, through the interactive exploration, they can figure out the underlying construction problems in regions, disclose the potential drawbacks of the current policies, and have a better construction management by having the knowledge of the expected cost of future construction requests. Third, for researchers are allowed to have a platform to investigate more complex urban dynamics through introducing heterogeneous urban information (e.g. traffic flow, check-in data, and user opinions and sentiments in social media) into our system. Our system can be accessed via the following link: http://junipertcy.info/urbcomp/index.html , and a snapshot of our system is shown in Figure 1.

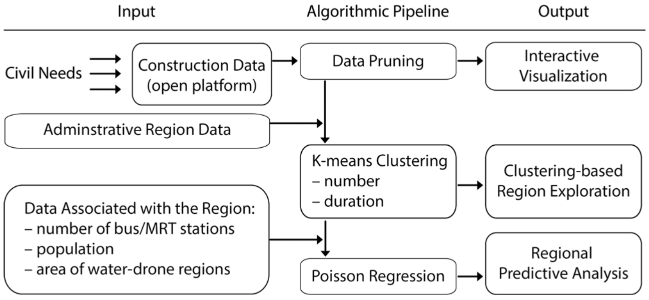

Input
Algorithmic Pipeline
Output
Civil Needs
Construction Data (open platform)
Data Pruning
Interactive Visualization
Adminstrative Region Data
K-means Clustering
– number
– duration
Clustering-based Region Exploration
Data Associated with the Region:
– number of bus/MRT stations
– population
– area of water-drone regions
Poisson Regression
Regional Predictive Analysis

Figure 2: System Framework.

2. SYSTEM FRAMEWORK

The system framework of this paper is shown in Figure 2. First, the construction request data collected from an open data platform is used as the input data of our system. The construction request data, which is daily updated, contains a collection of construction requests. Each construction record contains the reason, the geographical location, and the time and duration of a construction. In addition to the construction request data, we also obtain the data of administrative regions, i.e., villages, which are treated as the basic geographical units of our system. Each region/village is also associated with three regional information, including the numbers of Metro and bus stations, and the historical water-drone areas. In the following, we will briefly introduce the three components of our system, along with the corresponding outputs.

The first is to preprocess the raw data for providing a fundamental interactive data visualization and exploration system. We will show how to rule out improper data. In the system, users are allowed to specify (a) a particular year/month, (b) the time duration, (c) the type of construction requests, and/or (d) a specific region of interest, and then the corresponding data distribution as well as the geographical map visualization will be reported in a dynamic manner. Second, we group regions that share similar patterns of types of construction requests. An administrative village is considered as the basic unit of the geographical analysis, because its area, population, and functionality best represent "community" in general perception. Each construction request record is assigned to the corresponding village according to the geographical location. We perform K-means clustering algorithm on villages, and each output cluster contains a set of villages. Users are allowed to specify two feature options for clustering, (a) the number of construction requests in a region/village, and (b) the time duration of the construction. Third, we forecast the number of construction requests for a particular region of interest. We propose a generalized linear model to achieve the goal. For each region/village, the features are used to build the predictive model, including the time series of historical construction requests, the types of construction requests, the population data, the numbers of bus or Mass Rapid Transition stations, and the historical water-drone areas. We will report and analyze the error under the proposed predictive model.

3. DATA PREPROCESSING

The data used in this study comes from the *Data.Taipei*[4] platform, a governmental open data platform launched by the Taipei City Government in September 2011. Each public construction proposed by an associated governmental office or an executive company in Taipei City is required to apply through the Taipei City Office. After the application is approved, each construction will be asked to upload the GPS locations and pictures that reflect the daily construction work via a mobile app. All these construction application data were recorded and pushed to the *Data.Taipei* platform. We aim to analyze the dataset collected containing 19,882 records from June 23, 2014 to May 15, 2015. Each record includes the longitude and latitude of a location, an address, a date of initiation and accomplishment, a governmental office or company applying for the construction, and reasons of construction. Two records were deleted because of the abnormal coordinates, and 15 records were deleted because their reasons of constructions were absent. There were 19,865 records included in the analysis at last.

The 19,865 records were categorized to 14 construction types to characterize the major objective of construction and the urgency of construction, since the urgent repair of objects might reflect the different needs of the scheduled constructions. Categorization is input manually by the application department according to (1) reason of construction, which in the dataset were semi-categorized textual descriptions and (2) governmental office or company applying for the construction. The 14 construction types are listed in Table 1. In non-urgent constructions, most records (61.9%) were pipeline, powerline, and telecom cable related and few were associated with draining, road infrastructure or other developmental construction (11.68%). The proportion of "urgent

[4] http://data.taipei

repairmen" is 26.19%, and the "water pipeline" is the most common object.

Table 1: Types of construction requests.

Type	Major object of construction	# records
1	Gas pipeline	1111 (5.59%)
2	Powerline	3388 (17.06%)
3	Water pipeline	3660 (18.42%)
4	Telecom cable	4138 (20.83%)
5	Drainage infrastructure	738 (3.72%)
6	Road	773 (3.89%)
7	Other developmental construction	792 (3.99%)
8	Non-governmental construction	18 (0.09%)
9	Urgent repair of gas pipeline	936 (4.71%)
10	Urgent repair of powerline	65 (0.33%)
11	Urgent repair of water pipeline	3894 (19.60%)
12	Urgent repair of telecom cable	210 (1.06%)
13	Urgent repair of drainage infrastructure	93 (0.47%)
14	Urgent repair of other objects	4 (0.02%)

In addition to the construction categories, these records were also categorized to five groups according to the time duration of constructions. The five groups were "Very short" (less than 1 day), "Short" (1-2 days), "Median" (3-18 days), "Long" (19-76 days), and "Very long" (above 76 days). Each group contains 20.17%, 21.45%, 19.68%, 19.34%, and 19.36% of total data, respectively, and all proportions were approximating to 20%.

After data pruning, all records were allocated to 465 villages in the Taipei City by their locations. A demonstration website (http://junipertcy.info/urbcomp/index.html) was created to show the summary of duration and major objectives of construction and urgent repair with interactive visualization. Citizens could easily identify the details of construction work in which village they lived during the past year.

4. METHODOLOGY

4.1 Interactive Exploratory Visualization

We selected some dimensions from the pruned data for interactive visualization; specifically, the date the request was posted, the target village it took construction, the type and duration of the construction. These are the main variables for the user to get a full picture of the request data. To build a web-based site, these key dimensions were filtered with the dc.js[5] library for the front-end visualization. Many questions can be addressed and visualized using this tool, such as, what is the administrative village that has the most type of gas pipeline construction throughout the year? What about that within winter? What kind of constructions usually takes more time to complete? This tool could interest individuals or organizations who are concerned about the construction request data. Since the interactive visualization only shows some dimensions of the request pattern, we take a step further and ask whether these patterns could be similar in either adjacent or distant villages. For example, are there villages prone to have some same of their road infrastructures broken? What may the request pattern indicate?

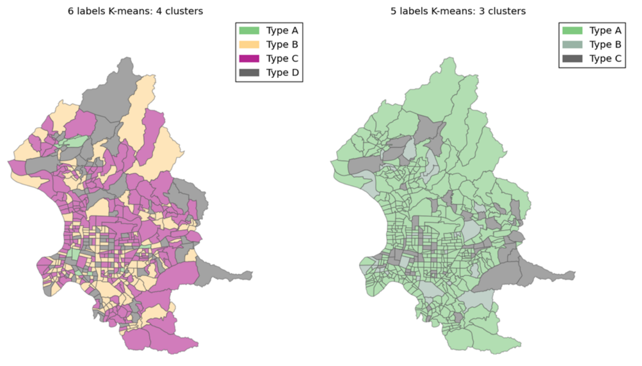
4.2 Clustering-based Region Exploration

In order to investigate the coherent hidden patterns of construction needs across different regions, we formulate the problem as finding clusters in the request number space. Specifically, we consider that the construction pattern of a region is characterized by various types of construction requests. Those regions sharing similar distribution of types of construction requests may either exhibit similar urban functions or suffer from similar urban construction problems. We use the clustering technique to group regions. Each region/village is represented by an m-dimensional vector, where m is the number of request types in the data. Each element in the vector is the number of a request type that is accumulated in the corresponding village.

We use the k-means clustering algorithm to find clusters of villages. Specifically, the number of needs over the 14 request types is considered as feature values for each village. There are 456 points representing villages to be clustered. A mined cluster contains a set of villages that share similar request patterns in terms of request types. To enable data exploration from clustering, our system allows two degrees of freedom in manipulating the analysis. The first one is the size of the region to be selected so that users are allowed to explore the construction dynamics under various geographical granularity levels. The second one is the type of request label to be clustered so that users will be able to target at only the request types of interest. By default, we choose the village as the minimum region of interest, and we labeled the data using the types of requests and the duration that the construction took.

We explored k from 4 to 6, and found k = 4 as the final choice because it best fits presentation and interpretation. As shown in Figure 3(a), the purpose of construction requests is clustered. We found some villages with similar request frequency patterns across all request types. For example, Type B expresses some villages are abundant in the request of powerline and telecom cable construction requests. Although there may be some environmental correlations among these villages, such as they are old-fashioned ones, lower altitude, or even sharing the same geological fault, we have no further clues to answer this issue. Similarly, we classified the duration of construction into five groups, as shown in Figure 3(b), the pattern is also clustered. It can be seen some villages are constantly having construction. All these coherent patterns can form interesting research questions to address in the future.

Although we did not include any geographic or demographic information in clustering, we see definite spatial patterns. This indicates that there are some geographical factors that affect the numbers of certain types of construction requests. To capture the characteristics, we apply the regression technique to model the relationship between the number of requests and the neighborhood factors, and obtain a formula to weekly predict the number of certain request types in a region in the future.



(a) Types of construction (b) time duration of construction

Figure 3: The results of the clustered regions/villages.

[5] https://dc-js.github.io/dc.js/

4.3 Regional Predictive Analysis

Next, we aim to build a model to predict the upcoming numbers of every construction request type in a region/village of interest. Finite distributed lag model was taken into account.[7] The lagged data are the data from one week, two weeks, and three weeks before the time interval of prediction. Since the clustering analysis revealed geographical regions which are dominant in certain request needs, suggesting these requests are affected by non-stochastic regional characteristics. For example, higher requests of road construction in some region/village might be associated with its regional characteristics, such as population, traffic situation or road condition. To capture those characteristics, we collected some other data from *Data.Taipei* platform depicting regional characteristics and potentially affect the number of certain request. The data include (1) number of bus stations, (2) number of MRT stations, (3) area of historic water-drone regions, (4) district population. These data were preprocessed using feature scaling technique by dividing each kind of regional features data according to its standard deviation.

We take advantage of the *Generalized Linear Poisson Regression* [8] to predict the future request number of each region/village. A generalized linear Poisson model relates a dependent variable, in our case, the amount of construction requests per request type, with a set of independent variables (i.e., time and the neighborhood characteristics) through an exponential function. To put these together in a formal form, by denoting $Y(t)$ as the number of construction requests for a particular construction type for week t in a given region/village, we can have the formulation of generalized linear Poisson model as follows:

$$Y(t) \sim \exp(\theta(t))$$

$$\theta(t) = \alpha + \varphi_1 \cdot \theta(t-1) + \varphi_2 \cdot \theta(t-2) + \varphi_3 \cdot \theta(t-3) + \hat{\beta} \cdot \hat{x}$$

where is the intercept or the mean value of theta, φ_1, φ_2, and φ_3 are autoregressive coefficients, $\hat{x}$ is a sequence of numerical *regional features*, namely: number of bus stations, number of MRT stations, area of historic water-drone regions, and district population. $\hat{\beta}$ is a sequence of weights, one for each of the above regional features.

Table 2: The parameter values learned in our model.

Parameter	Coefficient	p-value	Description
α	0.0543	0.000	Intercept or global mean
φ_1	0.0264	0.000	Number reports previous week
φ_2	-0.0025	0.743	Number reports two weeks before
φ_3	-0.0259	0.001	Number reports three weeks before
$\hat{\beta}_1$	0.1043	0.000	Standardized number of bus stations
$\hat{\beta}_2$	0.2185	0.000	Standardized number of Metro stations
$\hat{\beta}_3$	0.2610	0.000	Standardized number of population
$\hat{\beta}_4$	0.0472	0.000	Standardized area of water-drone regions

Fitting this model to our data means finding the values of each coefficient so that the relation expressed by the model is as close as possible to the observed data over a time span in a specific region/village. We take the Shih-lin district in Taipei as an example. The request type we want to model is the urgent repair of water pipeline data (label 11) over a 47 weeks-long time span. We use *Root Mean Square Error* (RMSE) of the testing data and predictive data as our evaluation metric. The definition of RMSE is given by: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, where y_i is the ground-truth value, $\hat{y}_i$ is the predicted value, and n is the number of instances. For time series data, the cross validation procedure is similar to the normal classification one, but the training set consists only of observations that occurred *prior* to the observation that forms the test set. Thus, no future observations can be used in the forecasting task. This procedure is sometimes known as *Rolling Forecasting Origin* [9]. The training data in the beginning set to the first 22 weeks and the rest are testing data. Then, we slightly move the timestamp forward to increase the training data, generate the predictive data, and calculate the RMSE by each time. The results of learned parameters obtained by fitting the training data in our model are presented in Table 2.

Table 3: RMSE values under various number of lagged weeks.

	$\theta(t-1)$	$\theta(t-2)$	$\theta(t-3)$	$\theta(t-1)+\theta(t-2)+\theta(t-3)$
RMSE	13.5092	11.0679	9.6868	3.4156

The absolute value of the coefficient is positively correlated with the effect of a variable on the number of requests. Among the regional features, we can see that historic flood area may have weakest impact on the construction requests. Both the numbers of MRT stations and village/region population have stronger effects on construction requests. To check which lagged time variable is more predictive to this model, we conducted only lagged one, only lagged two, and only lagged three variable separately into our regression model. We found that lagged-one variable is the most predictive parameter and the other two lagged variables are less deterministic with respect to the predictability of future numbers of construction requests. This might suggest the period or subperiod pattern of urgent repair of water pipeline request and will be different due to different request types and region/village we're modeling.

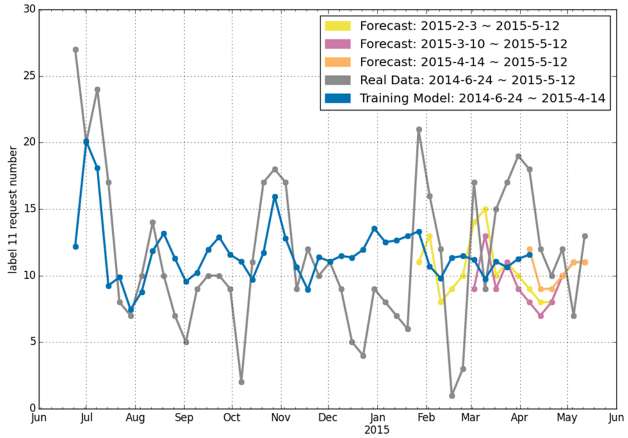


Figure 4: Curves of the numbers of construction requests over time: the forecasted, real data, and the fitted regression model.

Figure 4 and Figure 5 show the weekly plot of the numbers of construction requests for urgent repair of water pipeline from June 2014 to May 2015 and plot of the relation between RMSE and training weeks, respectively. The weekly line graph shows the patterns of real data, fitting model, and some predictive data. The plot of relation between RMSE and training weeks shows a sudden drop at about 40th week which is around March and April. To explain the phenomenon, we need to look into the weekly plot. We can see the pattern around March and April has successive increasing pulses which does not exist in the training data. This means that the model hasn't learned the pattern yet until the pattern is included into the training data. Therefore, the RMSE is higher before March and April, and then drops afterwards. Generally, if we have more data over much longer time span, which means there will be more patterns included in the training data, it is likely to generate a more accurate predictive model.

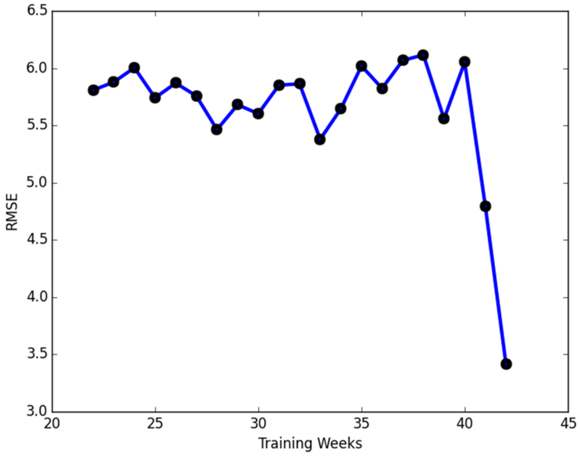


Figure 5: The RMSE values of the forecasted under our model.

5. RELATED WORK

Recent studies in the *Urban Computing* [10] area have exhibited the potential of visualizing and exploiting various urban data. *CityBeat* [1] used real-time geo-tagged photos to detect hyper-local trending activities. *LiveHoods* [2] employed check-in records to dynamically group and visualize locations with similar spatio-social characteristics. *UrbanSound* [3] classified urban sounds into a taxonomy to analyze urban events. *Urban Tribes* [4] detected the social intent of group photos to analyze the tribes in a city. *U-Air* [5] combined heterogeneous urban information, including road networks, place data, weather, and user mobility data, to infer the air quality values of location of interests in an urban area. In addition, *CityNoise* [6] utilize New York City 311 complaint data, together with road networks and user check-in data, to analyze the composition of noises in a particular location.

6. CONCLUSION

The system presented in this paper is general-purposed and can be applied to many other sources of data under the similar schema. The developed data exploration and request prediction model does not require strong constraints on the data being used; hence it is flexible and applicable for other urban data such as air quality values and urban noises. In the future, we will integrate the system with the data collection backend to provide a real-time online service that tracking and predicting various kinds of urban informatics. We believe it will be a powerful, informative and pedagogical analytic platform to understand urban dynamics, particularly for those expressing interactions among human activity, the city officials and the environment.

7. ACKNOWLEDGEMENTS

This work was sponsored by Ministry of Science and Technology of Taiwan under grant 104-2221-E-001-027-MY2. This work is also supported by Multidisciplinary Health Cloud Research Program: Technology Development and Application of Big Health Data, Academia Sinica, Taipei, Taiwan under grant MP10212-0318. In addition, the first author, Tzu-Chi Yen, especially thanks Qiao Lin and Xiaodong Liu for their fruitful discussions as well as the Sensoro team for the open research atmosphere.

8. REFERENCES

[1] C. Xia, R. Schwartz, K. Xie, A. Krebs, A. Langdon, J. Ting, and M. Naaman. Citybeat: Real-time Social Media Visualization of Hyper-local City Data. In *WWW* 2014.

[2] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *AAAI ICWSM* 2012.

[3] J. Salamon, C. Jacoby, and J. P. Bello. A Dataset and Taxonomy for Urban Sound Research. In *ACM Multimedia* 2014.

[4] A. Murillo, I. Kwak, L. Bourdev, D. Kriegman, and S. Belongie. Urban tribes: Analyzing Group Photos from a Social Perspective. In *IEEE CVPR* 2012.

[5] Y. Zheng, F. Liu, and H.-P. Hsieh. U-Air: When Urban Air Quality Inference Meets Big Data. In *ACM KDD* 2013.

[6] Y. Zheng, T. Liu, Y. Wang, Y. Liu, and Y. Zhu. Diagnosing New York City's Noises with Ubiquitous Data. In *ACM UbiComp* 2014.

[7] M. H. Pesaran and Y. Shin. An Autoregressive Distributed-lag Modelling Approach to Cointegration Analysis. *Econometric Society Monographs* 31: 371-413, 1998.

[8] R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson. Generalized Linear Models: with Applications in Engineering and the Sciences. *Wiley Series in Probability and Statistics*, 2012.

[9] R. J. Hyndman and G. Athanasopoulos. Forecasting: Principles and Practice. *OTexts*, 2014.

[10] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban Computing: Concepts, Methodologies, and Applications. *ACM Transactions on Intelligent Systems and Technology* (TIST), 5(3), 38, 2014.