Studying the Effects of Weather and Roadway Geometrics on Daily Accident Occurrence using a Multilayer Perceptron Model

Jeremiah Roland
College of Engineering and Computer Science, University of Tennessee at Chattanooga
Chattanooga, Tennessee
Jeremiah-Roland@mocs.utc.edu

Peter Way
College of Engineering and Computer Science, University of Tennessee at Chattanooga
Chattanooga, Tennessee
Peter-D-Way@mocs.utc.edu

Mina Sartipi
College of Engineering and Computer Science, University of Tennessee at Chattanooga
Chattanooga, Tennessee
Mina-Sartipi@utc.edu

ABSTRACT

One of the most common, yet dangerous, events that people face each day is driving. From unpredictable weather to hazardous roadways, there is a seemingly endless number of factors at play that can lead to vehicular accidents. Therefore, attempting to predict these accidents is a timely topic in today's research spectrum. The data used in this research consists of historical accident records from Hamilton County, Tennessee beginning in 2016 and continues to be updated daily, as well as the associated weather occurrences and roadway geometrics. To enhance heterogeneity a procedure was performed that generated non-accident traffic data based on our actual traffic accident data. This procedure is called negative sampling. These different data sets were combined and placed through a Multilayer Perceptron (MLP) machine learning model. The end results displayed a high collective correlation between accident occurrence and the various features considered in our proposed model, allowing us to predict with 77.5% accuracy where and when an accident will occur.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; *Statistical relational learning*; *Model verification and validation*;

KEYWORDS

Machine Learning, MultiLayer Perceptron Modeling, Predictive Analysis, Accident Prediction, Roadway Safety

ACM Reference format:
Jeremiah Roland, Peter Way, and Mina Sartipi. 2019. Studying the Effects of Weather and Roadway Geometrics on Daily Accident Occurrence using a Multilayer Perceptron Model . In *Proceedings of 14th International Science of Smart City Operations and Platforms Engineering , Montreal, QC, Canada, April 15, 2019 (SCOPE'19),* 6 pages.
https://doi.org/10.1145/3313237.3313304

1 INTRODUCTION

Research has found that multi-lane roadways are among the highest listings of most hazardous locations in our roadway networks [1]. The Highway Safety Manual (HSM) contains safety performance functions (SPF) for multi-lane roadways, estimated using data from specific states [6]. However, given the vast differences of traffic, environmental conditions, and crash related factors across our nation, there are few available and reliable prediction methods in use today. Within the state of Tennessee, there are many significant variations of the previously mentioned factors due to spatial and temporal differences across the area. This leaves a need for additional research into the jurisdiction's SPF needs, more accurately, a model to take into consideration spatial and temporal specificities of Tennessee roadways. This was achieved by providing a detailed analysis and predictive model of the various effects that spatial and temporal conditions (e.g. weather and roadway geometrics) can have on driving.

In this paper, the analysis and prevention of accident occurrence is explored via the use of machine learning techniques. Previous contributions to the preventive studies of accidents are explored in Section 2. Section 3 introduces and explains data used, as well as methods utilized in the testing and construction of a Multilayer Perception neural network. Section 4 analyzes results and walks through the architecture of the created MLP. Section 5 continues with a summary of the research completed, and suggests further work, as well as limitations encountered, and ends with acknowledgements.

2 RELATED WORKS

The effects of weather conditions on daily crash counts were analyzed using a discrete time-series model. In the project, an integer autoregressive model was introduced for modeling count data with time interdependencies. Then the model was built from daily car crash data, meteorological data, and traffic exposure data from three cities in the Netherlands: Dordrecht, Utrecht, and Haarlemmermeer [4]. Daily vehicle counts were collected for each road segment of the major road networks based on loop detector data. From this, day-to-day total amount of vehicle kilometers driven were calculated on the major road network of each city region. Weather data was also collected for the three cities and deaggregated into specific weather instances. For example, precipitation was broken up into precipitation duration, daily precipitation amount, rain intensity, etc. This type of deaggregation was done for wind, temperature, sunshine, precipitation, air pressure, and visibility. With all data

SCOPE'19, April 15, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6703-5/19/04…$15.00
https://doi.org/10.1145/3313237.3313304

then collected, an Integer-Valued Autoregressive model was used to find the significance of different weather occurrences on accident occurrence. It was discovered that several weather variables were significant in relation to accident occurrence.

State-specific Safety Performance Functions (SPFs) for rural interstates and rural 2-lane roads was used by [5] to identify the 20 segments of each type with the highest Potential for Crash Reduction (PCR). A Cost Benefit Analysis (CBA) was then performed, using appropriate Crash Modification Factors (CMF) for the types of crashes occurring in an effort to make an index that normalized the safety benefit of all roadway classes based on implementation cost. Model Minimum Uniform Crash Criteria (MMUCC) was used by [5] along with Knee Airbag Deployment models for identifying and classifying accident data. Once road segments with the highest PCR values were identified, CBA was used to identify which sites would provide a return on investment and in ranking the segments deserving treatment.

The effects of traffic and weather characteristics on road safety were examined by [9]. Gaps were identified, and needs for future research discussed as well. Roadway data such as average daily traffic, road density ratio, and speed were considered, as well as weather data such as precipitation, fog, and sunshine. It was found by [9] that Logit models were the dominant means of analysis used for crash severity, and time series analyses were the dominant means of analysis used for weather data.

A case study was conducted by [11] on predicting traffic accidents by utilizing and comparing the results of four different classification models of prediction. Those methods were linear Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Deep Neural Networks (DNN). In this study, a method of generating non-accident data was performed and called negative sampling. For each positive example (accident), the value of only one feature was changed among hour, day, and road ID, then the resulting sample was checked for a positive (match found) or negative (no match found) result. In the end, results dictated that the most optimal model was DNN.

3 RESEARCH METHODOLOGY

Our contribution to the study of vehicular accidents is partially owed to the unique metropolitan area of Hamilton County. A region with such diverse weather conditions over a comparably small area, coupled with Chattanooga Department of Transportation's database of roadway geometrics produces a perfect test bed for accident research. Research was completed in three distinct phases of work, reflected in Figure 1. The first phase, Data Collection, involved acquisition of all available data for research usage. The second, Data Pre-Processing and Aggregation, encompassed data cleaning, aggregation, and deaggregation as needed. The final phase of research, Modeling and Visualization, completed the process by providing insight into the complex situation of vehicular accidents.

3.1 Data Collection

- **911 Factors** include data taken from the Hamilton County Emergency Communications District. This was the beginning step in collecting data for the model. These call records initially included the physical address of the accident, the city the accident occurred within, the latitude/longitude coordinates, level of injury severity, as well as the time the accident was reported and the time it was resolved.
- **DarkSky Weather Factors** were obtained using the location and time information from the 911 data, allowing for weather conditions during the accident to be determined. DarkSky is an API for Python, which collects weather data from several weather resources and returns the data best suited to the specific location supplied, providing a definite spatial match. The specific weather conditions included in the reports were *Conditions*, *Temperature*, *Humidity*, *Visibility*, and more.
- **E-Trims Factors** include roadway geometric data taken from the E-Trims database. E-Trims is a massive database of roadway descriptive data used by the Tennessee Department of Transportation, containing information with regards to several different roadway factors for the roadways of Tennessee. The latitude and longitude locations for the 911 calls were utilized to locate roadway matches for locations in the E-Trims database. By connecting the specific 911 accident to the appropriate roadway, particular road geometrics that have correlations to traffic accident occurrences may be studied, such as *Pavement Width*, *Pavement Type*, *Land Use*, and more.

The complete details of utilized datasets are explained below, in Table 1.

3.2 Data Pre-Processing and Aggregation

The process of data aggregation combined information from our heterogeneous sources into one report. This constructed a full view of the circumstances surrounding an individual accident.

The 911 data was cleaned by removing *Fixed Time-Call Closed* and formatting the *Response Date* into separate *Time* and *Date* variables. Further, *Problem* was simplified to clearly state what the problem of the accident was. *Address* and *City* values were carried over without alteration, and *Latitude* and *Longitude* were adjusted to be properly formatted into floating point values. Next, any unnecessary columns were dropped, and necessary column values were added to hold our variables for the model. Duplicate values were then dropped from the data based on *Latitude*, *Longitude*, and *Problem* variables. This was accomplished by finding any call records with *Latitude* and *Longitude* values within .0001 of one another, then the call record with the highest *Problem* was retained.

Accident was considered the dependent variable, which displayed if the corresponding entry contains data for an accident or a non-accident (1 or 0, respectively). Initially, our dataset solely included entries for accidents which led to an imbalance in our model. We needed to generate non-accident data to act as a balancing agent so our model could predict accident and non-accident traffic data. To achieve this, non-accidents were gathered by negative sampling of actual accidents. To get these negative samples, a process outlined in [11] was followed. To summarize, 2 different aspects of data were adjusted: *Hour* and *Date*. For these two aspects, tests were ran to see if there were occurrences of accidents happening at the new hour or date. For example, hour of an entry was changed to a random hour between 0 and 23, inclusive, while excluding the entry's original hour. After the change, the new entry was compared to others to see if there was a match (if another accident happened at that hour on that day). If there was a match, then the sample was considered positive, and the next entry was used to repeat the process. However, if there was no match, then the adjusted entry was saved as a negative sample, indicating no records of an accident happening at the given time and date. This process was repeated for every entry in the dataset, yielding an overall increase of roughly 300% over the original number of entries. After the negative samples for hour were found, then negative samples for date were constructed. Similar to hour, the dates of each entry were randomly changed to any date between the beginning and end of the current year, excluding the day the entry was originally found.

Table 1: Data Features Used in Study

911 Variables	Explanation
Accident	No Accident(0) or Accident (1)
Latitude/Longitude *	GPS coordinates of record location
Date/Time *	Date and Time at which record occurred
Problem *	The level of injury occurred
Hour	The hour of the day record occurred
Month	Month of the Year record occurred
Weekday	Day of the Week record occurred
Weather Variables	**Explanation**
Event / Condition *	Present weather conditions
Event / Condition Before *	Weather conditions from previous hour
Rain Before	Presence of Rain in previous hour (0 or 1)
Temperature	Temperature at time of record
Temp Max / Min	Maximum / Minimum of the daily temperature
Dewpoint	Air temperature required for water vapor saturation
Humidity	Amount of water vapor in the air (0 to 1)
Cloud Coverage	Percentage of the sky covered by clouds (0 to 1)
Precipitation Intensity	Intensity of precipitation at time of record
Precip Intensity Max	Maximum level of precipitation for day of record
Road Variables	**Explanation**
Land Use	Use of surrounding area (Rural, Public Use, etc)
Route	DOT identifier for road
Log Mile	Exact location along route (miles)
Access Control	Access level citizens have to roadway
Operation	Pathway of roadway (1 or 2 way)
Pavement Type	Makeup of roadway surface (gravel, asphalt, etc)
Pavement Width	Width of roadway at record location (feet)
Thru Lanes	Lanes used for through traffic at record location
Num Lanes	Number of lanes on route at record location
Ad Sys	Road System (interstate, STP rural, etc)
Gov Cont	Government Control (private, state park, etc)
Func Class	Function Class (municipal highway agency, etc)

Note: Variables with * are used for data pre-processing, and are not used in the model

Event and *Condition* were taken from Dark Sky API and state the weather related conditions that were present during the 911 accident. Condition is more verbose than the event variable, though they both state the same information. These two were used to create the aggregated weather variables *Rain*, *Fog*, *Snow*, *Clear*, and *Cloudy*. These five aggregated variables represent the presence of their respective weather occurrences. *Event Before* and *Condition Before* represent the weather events from 1 hour before the 911 accident. These were used for the aggregated variable *Rain Before*, which indicates if there was rain during the previous hour.

Finally, once collection of accident data and negative samples was completed, all data was appended, and passed through the model for training and testing.

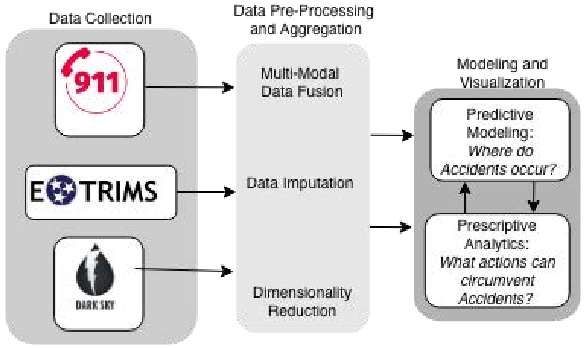
Data Collection
911
E-TRIMS
DARK SKY
Data Pre-Processing and Aggregation
Multi-Modal Data Fusion
Data Imputation
Dimensionality Reduction
Modeling and Visualization
Predictive Modeling: Where do Accidents occur?
Prescriptive Analytics: What actions can circumvent Accidents?

Figure 1: Illustration of Data Collection and Processing

3.3 Modeling and Visualization

Research has shown that different regression style models examine traffic flow differently, and as such, lead to varying results [9]. For example, previous research shows that Poisson distribution proved viable for use in accident frequency analysis related to modeling accident frequency. Poisson also proved superior to traditional linear regression in highway safety application usage [2]. Additionally, Negative Binomial models are plausible for use in exploration of crash severity, as shown in previous work [6]. Ordered logit/probit models are routinely applied as well, although their use highly depends on the number of levels involved with injury severity [6]. With cases of only two levels of injury severity, many previous research teams chose to utilize binary logistic modeling [7, 8, 10]. Finally, [3] utilized an ordered regression model, investigating five different levels of injury ranging from no injury to fatal.

Our team conducted several different types of testing when attempting to determine which analysis method would best fit our data and project. Some of these were regression techniques, while others were machine learning techniques. Their results were overall lackluster. Select K Best testing was performed before utilization of the MLP network testing various numbers of the existing variables for possible dimension reduction. However, once the results of the various Select K Best tests, with K ranging from 5 to 25, were compared to the standard results of the MLP the K Best tests under performed in both accuracy and area under the curve. These statistics were demonstrated by a range of 66.53-77.35% and 50-83.03%, respectively.

Additional tests were executed, and results continued to not perform to levels of the MLP network. Accuracy based off of Naive Bayes was a mere 62%, while the standard accuracy score test provided by Sklearn Metrics outperformed the Naive Bayes at 67.76%. However, both tests fell below MLP performance.

Bearing the aforementioned information in mind, an MLP Model was chosen for our project's machine learning technique, as we use labelled inputs for classification prediction, which MLPs are suitable for. Furthermore, MLPs are very flexible with the use of data, which is extremely beneficial to our project as our dataset is very complex and intricate. MLP networks are comprised of an input layer, an output layer, and at least one hidden layer between the two. The details of the architecture used by our model are displayed in Table 2. Compilation was provided initially by binary cross-entropy, which is particularly useful for binary results and classification. However, further investigation determined MSE (mean squared error) to be the superior option due to decreased loss even at lower epoch counts (300 versus 3000). This dramatic reduction in loss was at the cost of roughly 2% of accuracy The optimizer of the architecture was Nadam, an extension of Adam style optimization with Nesterov Momentum incorporation.

Table 2: MLP Neural Network Architecture

Layer	Location	Type	Node	Activation
1	Input	Dense	30	Sigmoid
2	Hidden	Dense	28	Sigmoid
3	Hidden	Dropout	-	Sigmoid
4	Hidden	Dense	20	Sigmoid
5	Hidden	Dense	18	Sigmoid
6	Hidden	Dense	10	Sigmoid
7	Hidden	Dropout	-	Sigmoid
8	Output	Dense	1	Sigmoid

In the case of sigmoid activation (shown in Equation 1), there is a bounded and differentiable real function whose definition is for all real input values with a non-negative derivative at each point.

$$S(x) = \frac{e^x}{e^x + 1} \tag{1}$$

After a comprehensive review of the analysis model, conclusions were made that the current data inputs and specific machine learning model used provided the best suited prediction accuracy currently available.

4 RESULTS

Figures 2 and 3 were constructed to examine the efficiency of the proposed MLP model. Figure 2 demonstrates the Receiver Operating Characteristic curve (ROC), reflecting the high diagnostic abilities of the binary classification system, with an Area Under the Curve (AUC) score of 81.37%. Figure 3 reflects the first twenty records in the dataset, with the actual accident status shown by circles and the predicted status shown by Xs. As shown in the figure, only three out of the twenty entries were inaccurately predicted.

Figure 4 demonstrates the accuracy and loss of the MLP model across the 3000 epochs used for testing. As evident in Figure 4, the accuracy of the training set was marginally higher than the testing set, with a final accuracy of 77.50% on training and 76.92% on testing. The loss finished at 0.1537% for training, while the testing loss was marginally higher, at 0.1602% on the last epoch. In each traffic accident, each individual factor considered will have varying degrees of involvement in the accident, from little or no involvement to playing a critical part in accident occurrence.

Hamilton County has provided a solid foundation and framework for this type of analysis and testing due to its varied topography and diverse landscape. Overall, this project has the unique potential to benefit all residents of Tennessee while also providing vast potential for expansion beyond its current scope. Drivers will benefit from better safety, peace of mind, and a reduction of car/insurance payments. Local governments and residents will benefit from the resulting gross reduction in total human and property damage as well as economic and medical costs. Furthermore, the methodologies explored here can be expanded or adjusted to fit additional districts of need.

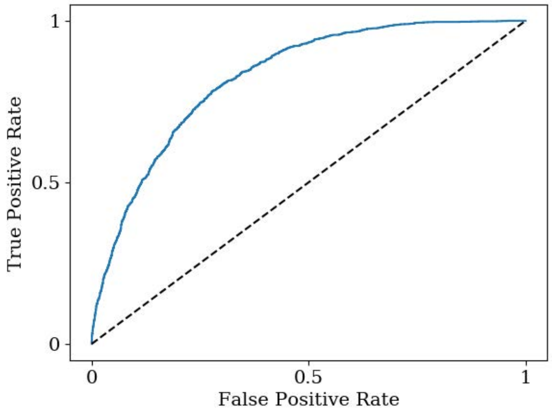

Figure 2: ROC curve of described parameters: AUC of 81.37%

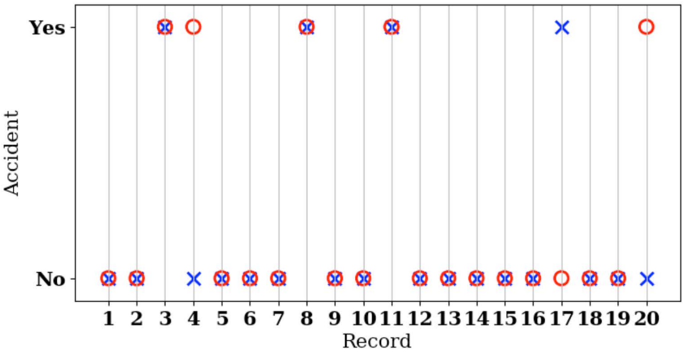

Figure 3: Demonstration of first twenty entries and prediction result

5 CONCLUSIONS

This paper investigated the potential correlations between traffic accident occurrence, weather occurrences, and roadway geometrics. When considered individually, the features used in our study reflected as being inconsequential when considering an accident occurrence. However, with the results from our machine learning reflecting a high accuracy in accident prediction, we concluded that many features are not only correlated with accident occurrence, but are also correlated with each other. With the diversity of our datasets used to accumulate appropriate data, we can conclusively and confidently state that our findings are sound, accurate, and can provide immeasurably important assistance with accident prediction, prevention, and relief.

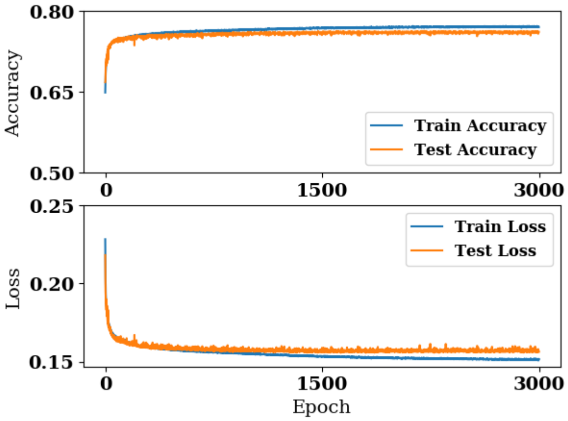

Figure 4: Training and testing accuracy (top) and loss (bottom) across the MLP 3000 epoch history

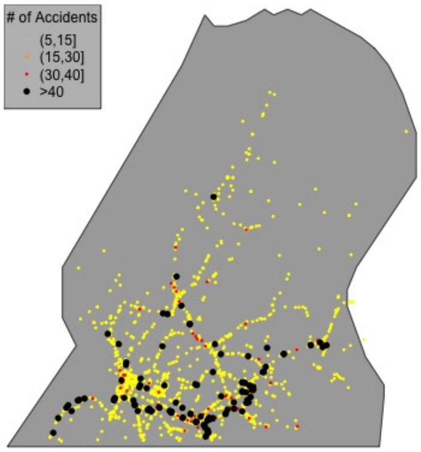

Figure 5: Concentration of Accidents in Hamilton County, Tennessee

5.1 Limitations

Our main limitations were the restrictions on our data. Some of the data retrieved from the E-Trims database had an unfortunately large amount of empty values, restricting their viability and use in the investigation. This lack of complete data was also a problem with some of the weather variables collected from DarkSky. Furthermore, with the inherent uncertainty and complexity of accidents, trying to take into consideration all of the potential factors at play during an accident is simply not plausible. A potential fix to these limitations is to seek out alternative sources of information relating to accidents.

5.2 Future Works

For potential future work regarding this topic, access to a higher quantity of data, such as driver specific data (e.g. age, gender, intoxication levels, etc.), additional roadway data (e.g. street illumination, more time-sensitive traffic volume counts, etc.), and more descriptive weather data (e.g. time to sunrise/set, days since last rained, etc.) would be beneficial. These additional factors would reveal even more insight into the correlations between traffic accidents and the various elements at play when an accident occurs.

As an example of future implementations, we are currently working on a project to utilize this accident prediction technique to inform travelers of potentially dangerous roadways. We plan to achieve this by circumventing the particular roadways that are deemed dangerous via an alternate routing algorithm.

5.3 Acknowledgements

We would like to thank the City of Chattanooga, Hamilton County Emergency Communications District, Tennessee Department of Transportation, Chattanooga Department of Transportation, and DarkSky for supplying data for this research. We also extend gratitude to the NSF US Ignite for allowing the opportunity for this project's existence through funding under award number 1647161.

REFERENCES

[1] AASHTO. 2010. An Introduction to the Highway Safety Manual. (2010). http://www.highwaysafetymanual.org/

[2] Mohamed A. Abdel-Aty and A. Essam Radwan. 2000. Modeling Traffic Accident Occurrence and Involvement. *Accident Analysis and Prevention* 32, 5 (2000). https://doi.org/10.1016/S0001-4575(99)00094-9

[3] S.R. Akepati and S. Dissanayake. 2011. Characteristics and Contributory Factors of Work Zone Crashes. (2011). Was a part of the transportation research board 90th annual meeting.

[4] Dimitris Karlis Brijs, Tom and Geert Wets. 2008. Studying the Effect of Weather Conditions on Daily Crash Counts Using a Discrete Time-Series Model. *Accident Analysis and Prevention* 40, 3 (2008), 1180–90. https://doi.org/10.1016/j.aap.2008.01.001

[5] N. Stamatiadis G. A. Winchester R. R. Souleyrette J.G. Pigman E. C. Davis, E.R. Green. 2015. Highway Safety Manual Methodologies and Benefit-Cost Analysis in Program-Level Segment Selection and Prioritization. (2015).

[6] Jun Liu Khattak, Asad and Meng Zhang. 2017. Highway Safety Manual: Enhancing the Work Zone Analysis Procedure Southeastern Transportation Center. (2017).

[7] Y. Li and Y. Bai. 2008. Development of Crash-Severity-Index Models for the Measurement of Work Zone Risk Levels. *Accident Analysis and Prevention.* 40, 5 (2008).

[8] C.F. See. 2008. Crash analysis of work zone lane closures with left-hand merge and downstream lane shift. (2008).

[9] Athanasios Theofilatos and George Yannis. 2014. A Review of the Effect of Traffic and Weather Characteristics on Road Safety. *Accident Analysis and Prevention* (2014). https://doi.org/10.1016/j.aap.2014.06.017

[10] J. Weng and Q. Meng. 2011. Analysis of driver casualty risk for different work zone types. *Accident Analysis and Prevention* 43, 5 (2011).

[11] Zhou Xun Yang Tianbao Tamerius James Yuan, Zhuoning and Ricardo Mantialla. 2017. Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study. In *In Proceedings of 6th International Workshop on Urban Computing*. Paparazzi Press, Halifax, Nova Scotia.