Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips

Juliana Freire Aline Bessa Fernando Chirigati Huy Vo Kai Zhao
New York University

Abstract

Traditionally, data cleaning has been performed as a pre-processing task: after all data are selected for a study (or application), they are cleaned and loaded into a database or data warehouse. In this paper, we argue that data cleaning should be an integral part of data exploration. Especially for complex, spatio-temporal data, it is only by exploring a dataset that one can discover which constraints should be checked. In addition, in many instances, seemingly erroneous data may actually reflect interesting features. Distinguishing a feature from a data quality issue requires detailed analyses which often includes bringing in new datasets. We present a series of case studies using the NYC taxi data that illustrate data cleaning challenges that arise for spatial-temporal urban data and suggest methodologies to address these challenges.

1 Introduction

Cities are the loci of resource consumption, of economic activity, and of innovation; they are also the cause of our looming sustainability problems and where those problems must be solved. Our increasing ability to collect, transmit, and store data, coupled with the growing trend towards openness [4, 14, 17, 23, 24, 34], creates an opportunity to leverage these data and make cities more productive, livable, equitable, and resilient. *Urban data* is unique in that it captures the behavior of the different components of a city, namely its citizens, existing infrastructure (physical and policies), the environment (e.g., weather), and interactions between these elements [18]. To understand a city and how its multiple elements interact, intricate analyses are necessary. As in any data analysis process, data cleaning is of crucial importance to bring data from a "messy" to a "tidy" state [40].

Data cleaning may be achieved through a multitude of methods, including filtering operations, statistical analysis, outlier detection, and missing value imputation. Traditionally, this is performed as a pre-processing step: a function $DirtyData \rightarrow CleanData$. We argue that data cleaning must be an integral part of the inherently iterative data analysis cycle: data cleaning must be applied on the fly. While exploring a new dataset, constraints that should be checked in the cleaning function, and which might not be evident at first, are naturally discovered. Consider Figure 8, which shows visualizations of NYC taxi trips on a map. These elicit the fact that the data contain pickups and dropoffs inside the rivers and in the ocean. Since there are no amphibious taxis, these represent erroneous data. This finding suggests the creation of a rule that checks whether the GPS coordinates for the trips are within polygons that lie on land.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

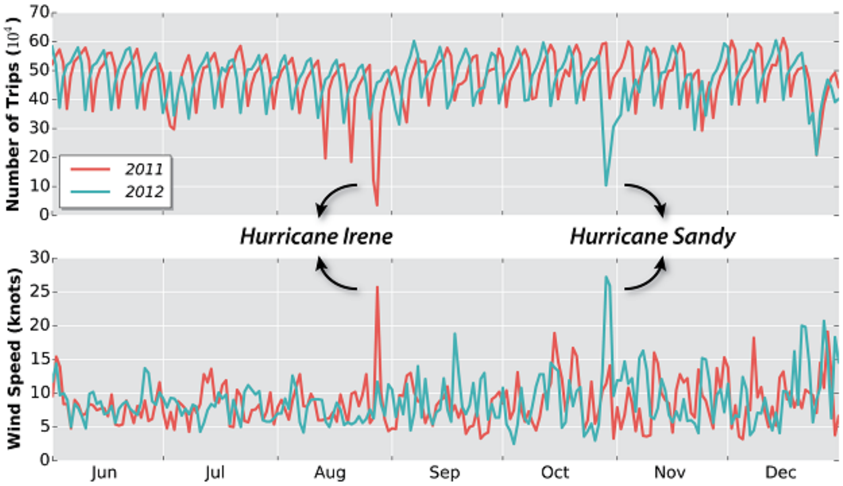


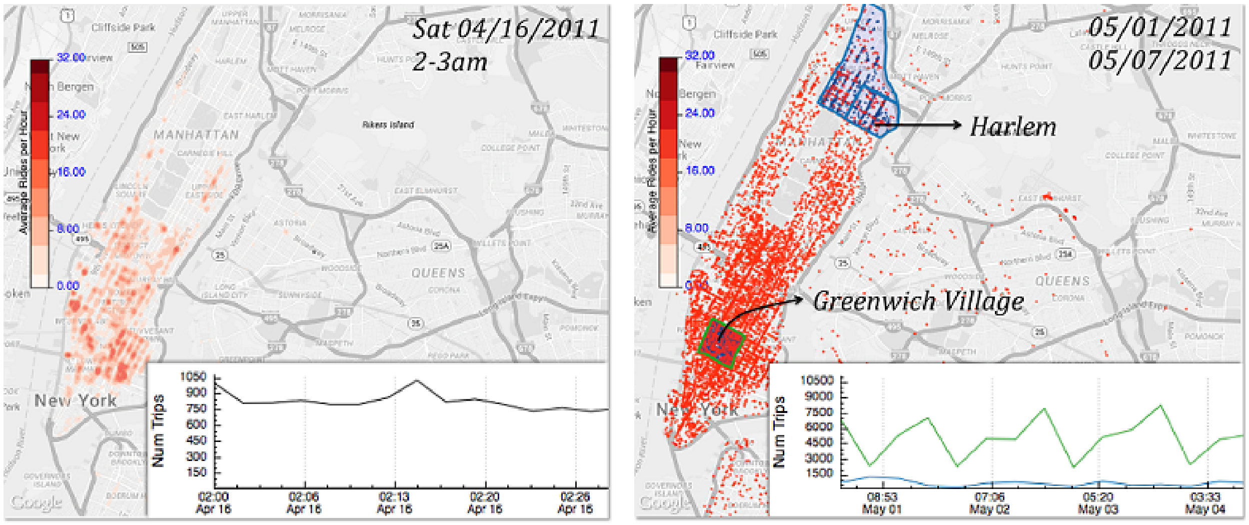
Figure 1: The plot on the top shows how the number of trips varies over 2011 and 2012. While the variation is similar for the two years, there are clear outliers, including large drops in August 2011 and in October 2012. However, examining the variation in wind speed during the same period (plot on the bottom), we can observe an inverse correlation: the large drops in the number of trips happened when the wind speeds were abnormally high. In fact, these correspond to two hurricanes: Irene and Sandy.

Besides the need to refine a cleaning function as the user gets more familiar with a dataset, different questions that arise during exploration may require different cleaning strategies. Thus, we need a function $DirtyData \times UserTask \to (CleanData, Explanation)$. For example, to create machine learning prediction models, cleaning steps are often applied to remove outliers from historical data. In contrast, to understand specific behaviors (or events), outliers are actually the central objects of the study.

As domain experts explore urban data corpora and consider different datasets, hypotheses are formulated and tested, and interactions among the different components of a city are untangled. In this process, when new datasets are brought into the investigation, seemingly erroneous data points identified when a dataset is analyzed in isolation may actually uncover features that explain important phenomena. Consider the top plot in Figure 1, which shows the number of daily taxi trips in New York City (NYC) during 2011 and 2012. While the distribution of trips over time is very similar across the two years, we observe large drops in August 2011 and October 2012. Standard cleaning techniques are likely to classify these drastic reductions as outliers that represent corrupted or incorrect data. However, by integrating taxi trips with the wind speed data (bottom plot in Figure 1), we discover that the drops occur on days with abnormally high wind speeds, suggesting a causal relation: the effect of extreme weather on the number of taxi trips in NYC. Removing such outliers would hide an important phenomenon.

Answering the question *"Is it dirt or a feature?"* is challenging because experts often need to go beyond a single dataset to seek explanations. Given the plethora of components interacting in urban environments, the integration possibilities for the data exhaust produced by these components are endless. As a point of reference, in the past two years, NYC has published over 1,300 datasets [24] and the city of Chicago has made available around 1,000 datasets [23], and these datasets represent just a small fraction of the data being collected by these cities [4].

In addition to the complex interactions among different datasets, the nature of urban data poses further challenges. First, metadata is limited or inexistent. Many of the datasets are derived from spreadsheets and contain incomplete schema information. Integrity constraints are not provided and type information often needs to be inferred [7]. Second, because urban data is often *spatio-temporal* [4, 7], there are many data slices to consider. For instance, NYC taxi trips are associated with GPS coordinates with time precision in seconds.



(a) (b)

Figure 2: (a) Heatmap of pickups on a Saturday uncovers popular nightspots. (b) Comparison between pickups in Harlem (blue time series) and Greenwich Village (green time series) shows that Harlem is underserved by taxis.

Consequently, such data can be aggregated into different spatial resolutions (e.g., neighborhoods, zip codes, and boroughs) and temporal resolutions (e.g., hourly, daily, weekly, and monthly) during the analysis process. Depending on the resolution, dirty data may become easily identifiable or completely hidden. For instance, missing data along an avenue over the course of an hour can be easily detected when looking at a finer scale (e.g., hourly and zip codes), whereas coarser resolutions (e.g., daily and boroughs) may hide these issues depending on the aggregation function used (see Figure 3). Spatio-temporal patterns present additional challenges to cleaning: data that may be detected as dirty in a specific time period or spatial region may in fact constitute a pattern in space and time. For example, as shown in Figure 1, there are significant drops in the number of trips on Christmas and New Year's day: these are not dirty data but recurring, yearly patterns.

The complexity of urban data coupled with the sheer number of available datasets and their numerous possible interactions make it hard to pinpoint what is an error and what is a feature. In this paper, we discuss challenges involved in cleaning spatio-temporal urban data. We use the New York City Taxi data obtained through a FOIL request[1] and present a series of case studies that illustrate different problems that arise. We also describe techniques that can aid in exploratory data cleaning and outline directions for future research.

2 The New York City Taxi Data

In New York City, every day there are over 500,000 taxi trips serving roughly 600,000 people [6]. Through the meters installed in each vehicle, the Taxi & Limousine Commission (TLC) captures detailed information about trips. Each trip consists of two spatial attributes (GPS readings for pickup and dropoff locations), two temporal attributes (pickup and dropoff times), and additional attributes including taxi identifier, distance traveled, fare, medallion code, and tip amount.

Taxis are unsuspecting sensors of urban life and their data exhaust can help uncover characteristics of the city that are of economic and social importance [12]. For example, by examining patterns involving taxi pickups and dropoffs in NYC at different times, we can discover popular destinations (e.g., popular night spots) and neighborhoods that are underserved by taxis (see Figure 2). In addition, we can discover various events (or

[1] A new version of this dataset was recently released by the Taxi & Limousine Commission [37].

exceptions), such as road closures and hurricanes, and their effect on traffic [10]. It is also possible to identify functional regions, such as tourist attractions, shopping centers, workplaces, and residential places, based on urban human mobility patterns [27, 41].

Taxi data are also valuable in that they can be used to derive other types of data. For example, traffic speed and direction information can be derived from taxi data, helping deal with the sparsity of speed sensors which cover only a limited number of road segments in Manhattan [26]. Similarly, the concentration of $\mathrm{PM}_{2.5}$ in a city, which is a metric for air quality, can be inferred from traffic flow in locations where there is a limited number of air-quality monitor stations [43]. Another important use for these data is in the analysis of what-if scenarios. Savage and Vo [31] showed that, even though NYC residents are dissatisfied with taxi availability during rush hour, increasing the number of taxis in this period would lead to congestion and a significant reduction in traffic speed. Ota et al. [25] developed a real-time, data-driven simulation framework that uses historical trip data to supports the efficient analysis of taxi ride sharing scenarios.

2.1 A Quick Look into the Taxi Data

Given the wide range of applications that are enabled by the taxi data, understanding its quality is of utmost importance. Here, we focus on the taxi datasets collected from 2008 to 2012. Table 1 summarizes a few statistics associated with these datasets. Note that, while fare information is available for trips paid either by cash or by credit card, tip information is only available for trips paid by credit card and, as a consequence, reported tip statistics ("Tip Amount" column in Table 1) refer exclusively to credit card trips.

The averages reported show that taxi trips often last less than 17 minutes, cover less than 7 miles, and cost about US$10.00. But there are exceptions which represent potential data quality issues. We computed the average fare values for trips that were exclusively paid by credit cards: US$12.56 for 2008, US$10.81 for 2009, US$11.12 for 2010, US$11.20 for 2011, and US$11.84 for 2012. By crossing these values with tip values on Table 1, we observe that, for most years, passengers who paid with credit cards gave tips of about 20% of the corresponding fare amounts, which is consistent with the customary tip values for other services in New York City. In contrast, in 2008 and 2009, the average tip amount is much lower: 0.7% and 3.5%, respectively. This may have been an error in the way the data were reported. The issue seems to have been resolved by the end of 2009, when the average credit card tip increased to US$2.04.

If we turn our attention to trip duration, distance, and fares, there are clear discrepancies: average trip durations were significantly smaller for 2009 and 2010, even though their corresponding average distances were roughly twice as long as those of 2011 and 2012; the average trip duration for 2008 was significantly higher than those recorded for 2011 and 2012; the average fare in 2009 is much lower than in other years; and the average fare for 2008, US$0.09, is too low.[2] These suggest quality issues in the data that require further investigation.

The table also shows the presence of invalid, negative values in 2010. These are clearly wrong and should be removed during cleaning: the negative values do not carry useful semantics—there is no such thing as negative miles. On the other hand, the decision is not so clear cut for other, positive values. An example is the tip of US$938.02 (maximum credit card tip value for the 2010 dataset). While this could be an error in data acquisition or in the credit card information, it could also be the case that a wealthy passenger overtipped the taxi driver (e.g., a financier that had just made a big profit). Given that negative values are only found in 2010, it may be the case that data for the other years were cleaned and had negative values removed, or that some improvement was made in the way data was transferred from taxis to servers. This underscores the importance of including provenance information [13], detailing the cleaning operations applied to released datasets. Such provenance is essential to ensure that analyses across the different datasets are consistent.

[2] Since the average fare for credit card trips only in 2008 is US$12.56, the low overall average may indicate that cash payments were not properly reported by drivers.

Dataset	Statistic	Trip Duration (min)	Trip Distance (mi)	Fare Amount (US$)	Tip Amount (US$)
2008	Min	0.00	0.00	0.00	0.00
	Avg	16.74	2.71	0.09	0.10
	Max	1440.00	50.00	10.00	8.75
2009	Min	0.00	0.00	2.50	0.00
	Avg	7.75	6.22	6.04	0.38
	Max	180.00	180.00	200.00	200.00
2010	Min	-1,760.00	-21,474,834.00	-21,474,808.00	-1,677,720.10
	Avg	6.76	5.89	9.84	2.11
	Max	1,322.00	16,201,631.40	93,960.07	938.02
2011	Min	0.00	0.00	2.50	0.00
	Avg	12.35	2.80	10.25	2.22
	Max	180.00	100.00	500.00	200.00
2012	Min	0.00	0.00	2.50	0.00
	Avg	12.32	2.88	10.96	2.32
	Max	180.00	100.00	500.00	200.00

Table 1: Statistics for the taxi datasets. Tip amount is available for trips paid by credit card only.

2.2 Exploring Quality Issues in Spatio-Temporal Data

Computing simple statistics over attributes can help uncover potential issues in a dataset. However, in the case of taxi trips, substantial complexity is added to the cleaning process due to the spatio-temporal nature of the data. Manual (exhaustive) exploration is time-consuming and, for large datasets such as the taxi data, it is impractical. For example, temporal aggregation of a years worth of data into a discrete set of hourly intervals results in over 8,000 data slices to be explored.

Recently, techniques and systems have been proposed to streamline and better support exploratory analyses of spatio-temporal data. These include visualization and interaction techniques that allow users to freely explore the data at various levels of aggregation [2, 12, 35, 39] as well as indexing strategies that speed up the computationally expensive point-in-polygon queries required for this type of data [11]. However, effective interaction with spatio-temporal visualizations remains a challenge [15, 28] and, even by using these techniques, domain experts may still need to examine a prohibitively large number of spatio-temporal slices to discover interesting patterns and irregular behaviors, including potential errors in the data. As a step towards addressing this problem, we proposed a scalable technique to automatically discover spatio-temporal events and guide users towards potentially interesting data slices [10] (see Section 3.1 for details). Note that mining for exceptions at different levels of aggregations for relational data has been studied before in the context of OLAP data cubes [29, 30].

While automatic event detection can help steer users to interesting data slices, the user is still faced with the challenge of understanding the events and determining whether they correspond to data quality issues or important features. In [8], we presented the *Data Polygamy* framework, which enables the discovery of relationships between spatio-temporal datasets through their respective events. These relationships provide hints that can help explain the events. The relationship between the number of taxi trips over time and wind speed shown in Figure 1 is one example of a relationship discovered by the Data Polygamy framework.

Techniques that enable users to interactive explore spatio-temporal data, support automatic event detection, and aid in the discovery of relationships among disparate datasets are essential in the discovery (and resolution) of potential data quality issues in spatio-temporal data. In what follows, we present a series of case studies that show how these techniques can help users identify and reason about quality issues in spatio-temporal data.

Case Study	Possible Methodologies
Unusual Spatio-Temporal Behavior	Exploring Different Data Slices Exploring Different Data Resolutions Visualization Event Detection
Taxi Trips and Weather	Combining Multiple Datasets Visualization
Missing Data or Sparsity	Exploring Different Data Slices Exploring Different Data Resolutions Domain Knowledge
Taxis as Sensors	Exploring Different Data Slices Exploring Different Data Resolutions
Speed Computation	Exploring Different Data Slices Exploring Different Data Resolutions Outlier Detection Domain Knowledge
GPS Inaccuracy	Visualization Clustering
Ghost Trips	Domain Knowledge

Table 2: Case studies and possible methodologies for data cleaning.

3 Is It Dirt or a Feature?

In this section, we present case studies that showcase challenges involved in identifying potential quality issues in the NYC taxi data. These case studies demonstrate the importance of exploration and event detection in order to clean data. Table 2 summarizes possible methodologies to address the challenges that arise in each case study. It is worth noting that the decision of whether and how to clean the data depends on the application; such decision is outside the scope of this paper.

3.1 Identifying Unusual Behavior in Spatio-Temporal Data Slices

Consider Figure 3(a), which shows visualizations of the taxi data over four hourly intervals from 7am to 11am, on May 1st, 2011. Note that between 8 and 10am, there are virtually no taxis on 6th avenue between Midtown and Downtown. This anomaly was originally discovered by a user while visually exploring the data using the open-source TaxiVis system [12, 36]. In this case, the anomaly can be easily explained: 6th avenue was closed for the annual NYC 5 Boro Bike Tour.[3] In general, finding explanations for such events is challenging and may require the integration of multiple datasets [5, 8, 9].

To reduce the number of data slices the user has to consider, the usual approach is to apply different types of aggregation and produce visual summaries [1, 20]. These lead to a trade-off between the level of aggregation and the number of data slices to be explored. The use of a coarse (spatial or temporal) aggregation reduces the number of data slices, but it may result in loss of information. If we aggregate the data over time or spatially, as illustrated in Figures 3(b) and 3(c), the Bike Tour event would go unnoticed. Therefore, to detect events and dirty data in urban datasets, data must be analyzed at different granularities. Nevertheless, this is a hard and time-consuming task since it requires the examination of a prohibitively large number of spatio-temporal data slices. As a consequence, methods that automatically discover such events and guide users towards data slices that display interesting events are crucial for data cleaning.

As discussed in Section 2.2, we proposed an approach for automatically detecting events in spatio-temporal data [10]. Event detection is accomplished through the application of topological analysis on a time-varying scalar function derived from the urban data. We use the minima and maxima of a given function to represent

[3] `http://www.nycbikemaps.com/spokes/five-boro-bike-tour-sunday-may-1st-2011`

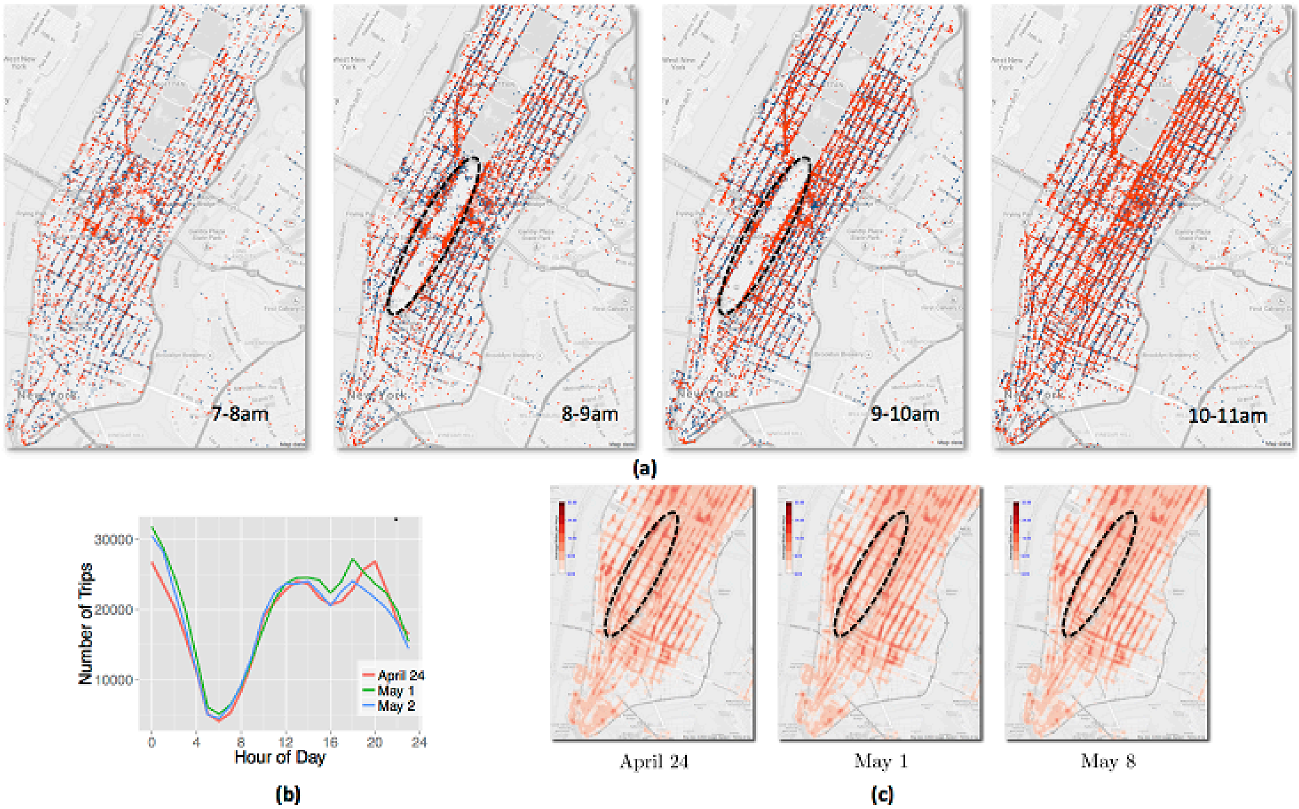


Figure 3: (a) Pickups (blue) and dropoffs (red) in Manhattan on May 1st from 7am to 11am. Notice that from 8am to 10am, there are virtually no trips along 6th Avenue. This avenue was closed to traffic at this time for the annual NYC 5 Boro Bike Tour. (b) The time series plots compare the number of trips that occurred in Manhattan on three Sundays in 2011: 24 April, 1 May, 8 May. It is difficult to distinguish between the three Sundays by using just the number of trips, even though an entire stretch of streets are blocked to traffic on May 1st. (c) The trips are aggregated over time and displayed as a heat map for the three Sundays. Note that the path of the bike tour (highlighted) looks similar in all heat maps.

the events in the data. Intuitively, a minimum (maximum) captures a feature corresponding to a valley (peak) of the data. For example, the lack of taxis along 6th avenue during the bike tour event forms a local minimum and is therefore captured using our technique. The use of topology also allows the detection of events that have an arbitrary spatial structure. In order to support a potentially large number of events, we designed an indexing scheme that groups similar patterns across time slices, thereby allowing for identification of not only periodic events (hourly, daily, and weekly events), but also of events with varying frequency (regular and irregular). Thus, unlike previous approaches that impose a rigid definition of what constitutes an event [3], our technique is flexible and able to capture a wide range of spatio-temporal events. Compared to techniques based on statistical analysis that support different kinds of events, our approach is computationally efficient and scales to large datasets. We also implemented a visual interface designed to aid in event-guided exploration of urban data that integrates event detection and indexing techniques. The interface allows users to interactively query and visualize interesting patterns in the data. We showed that experts applying this framework were able to quickly explain some of the events we found, but they were surprised by others which indicated potential problems they had to investigate. This suggests that our approach, as well as other approaches for event detection over spatio-temporal data, can be very useful for cleaning tasks.

The problem of event detection has been studied by the statistics and machine learning communities [16, 21, 22, 38]. However, the majority of the literature has focused on either purely spatial data or has accounted

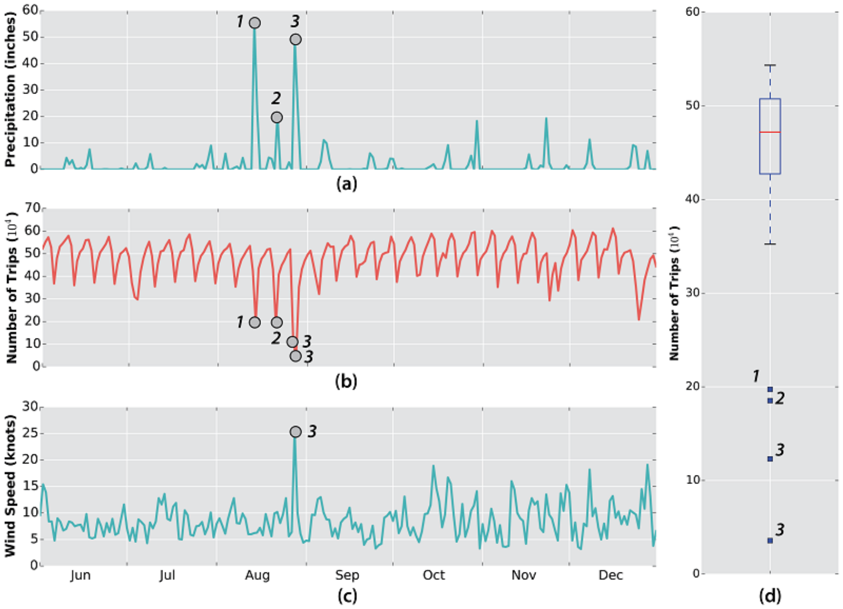


Figure 4: (a) Precipitation, (b) number of taxi trips, and (c) wind speed in a daily basis over the course of 2011. Interesting anomalies, which are detected when using a box plot with taxi data from August (d), are highlighted above, including heavy rainfalls (marked as 1 and 2) and hurricane Irene (marked as 3).

for temporal variations and effects via simplistic approaches such as exponentially weighted linear regression or data partitioning based on day-of-week or season. Furthermore, the time complexity for these approaches is exponential $O(2^N)$ in the number of *pre-defined* space-time partitions. In contrast, the topology-based method has polynomial time complexity [10].

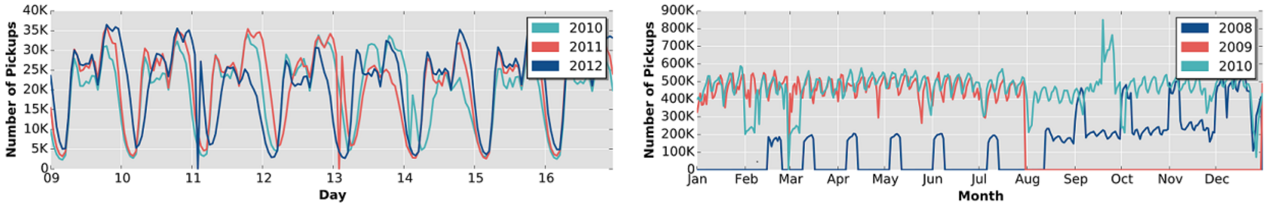
3.2 Combining Multiple Datasets: Taxi Trips and Weather

Consider the plot depicted in Figure 4(b), which shows how the number of taxi trips per day varies over 2011. Note the regularity in the trip distribution over time: the number of trips peaks on Fridays, and bottoms out on Sundays. There are a few exceptions when large drops are observed. Some of these drops can be easily explained, for example, on New Year's Eve and Christmas, which happen every year (see Figure 1). Others, such as the drops in August (labeled 1, 2, and 3), are not clear. By using standard statistical methods such as box plots (Figure 4(d)), these are detected as outliers. Since these points are anomalies, scientists could hypothesize that the data were corrupted on those days (e.g., loss of data when transferring information to servers) and classify these as dirt, removing them from the dataset.

However, further analysis of precipitation (Figure 4(a)) and wind data (Figure 4(c)) shows that these points correspond to different weather events that affected traffic in NYC, including heavy rains (anomalies 1 and 2) and hurricane Irene (anomaly 3). Therefore, anomalies in taxi data are not necessarily a product of dirty data. In these cases, they actually reveal interesting phenomena that demand further exploration.

The question of whether anomalies correspond to dirty data or features, as this example illustrates, may require one to look outside the data and bring additional datasets to the data exploration process. This is a challenging task, especially in the urban data context where a plethora of datasets is available: identifying meaningful connections among thousands of possible datasets is difficult and time-consuming.

Data Polygamy [8] is a scalable topology-based framework that allows users to query for statistically significant relationships and connections across thousands of spatio-temporal datasets. This is accomplished in three steps: (1) each attribute of the two datasets is transformed into a *scalar function*; (2) a topological data



(a)

(b)

Figure 5: (a) Comparison between the number of trips in March for different years: on two days, no trips are recorded at 2am. (b) Missing trips observed between 2008 and 2010, and an unusually number of trips at midnight in October 2010.

structure is computed for every scalar function which provides an abstract representation of the peaks and valleys and serves as an index to efficiently identify events; and (3) possible relationships are identified based on the similarity of the events from different scalar functions, and relationships that are statistically significant are returned. The framework not only drastically reduces the number of relationships one needs to analyze, but also uncovers surprising relationships that aids the data analysis process. While Data Polygamy provides insights into which data points are erroneous and which represent important features (or events), other methodologies can also help with this task. For instance, visualizations of the data may elicit connections across datasets (e.g., plots as depicted in Figure 4).

3.3 Missing Data or Sparsity?

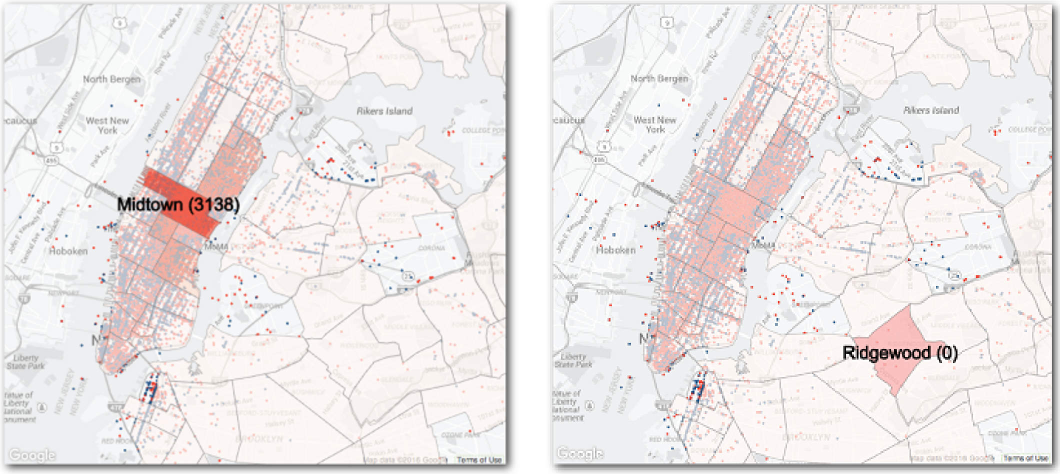
By examining the month of March for different years at an hourly resolution (Figure 5(a)), we observe that there are no trips at 2am on March 13th, 2011 and March 11th, 2012. Also, consider Figure 5(b), which depicts the trip distribution over years between 2008 and 2010. In 2008, we see several periods of missing data, and in 2009, there are no trips between August and December. In 2010, there was a week when the taxi pickups spiked up abnormally: there were 50,000 pickups during one hour at midnight on September 19th, 2010, whereas under normal conditions there are around 10,000 trips per hour. The challenge in this case is to identify whether these situations are due to missing data, data sparsity, or special events.

In Section 3.2, the abnormally small number of trips corresponded to days with extreme weather events. In contrast, the anomalies in Figure 5 are instances of dirty data. The drops in Figure 5(a) are likely due to inconsistencies in how the TLC dealt with Day Light Savings Time, while further examination of the data in Figure 5(b) showed that there is an unusually large number of consecutive and extremely short trips (lasting less than a minute), which cannot happen in practice, indicating that there was an error in data acquisition.

To uncover such issues, we had to first explore different data slices at different granularities: for instance, if data were aggregated by day, the missing data at 2am could not be identified. In addition, the large number of taxi trips observed in Figure 5(b) could be explained by using domain knowledge: we know that taxi trips cannot last less than a minute. This is a data cleaning rule, or a data constraint, that was uncovered *during* data exploration. This shows that the importance of having data cleaning be an integral part of data exploration.

3.4 Coverage: Taxis as Sensors

As discussed in Section 2, many applications have used taxis as sensors to infer new data, including air quality [43] and traffic speed and direction [26]. The quality of the derived data is highly dependent on how much of the city is covered by taxis. The *coverage* of taxis in a city represents the percentage of roads, neighborhoods, or boroughs that is visited by at least one taxi during an hour, a day, or a month. Data for a region is only recorded if that region is visited by a taxi.



(a)

(b)

Figure 6: The number of trips in (a) Midtown (Manhattan) and (b) Ridgewood (Brooklyn) on May 2nd (Monday) from 8am to 9am in 2011.

Taxi coverage in a city is often biased. Figures 6(a) and 6(b) depict the number of trips starting in Midtown (Manhattan) and Ridgewood (Brooklyn), respectively, on May 2nd from 8am to 9am in 2011. During that hour (peak time), around 20,690 trips are recorded by over 13,000 taxis in the entire city: while there are 3,138 trips with pickups in Midtown (Figure 6(a)), there are no trips leaving Ridgewood (Figure 6(b)). Out of the 132 neighborhoods in NYC, only 68 neighborhoods are covered by at least one taxi during that hour. The coverage of yellow taxis for NYC is around 51.50%, which means that nearly half of the city is not visited by any taxi. The coverage analysis reveals an instance of *data sparsity*: there is no data for several spatial regions, while there is too much data for others.

Depending on the task, the lack of coverage may severely limit the analysis. For instance, if a domain expert wants to build a human mobility model based on this dataset, a very detailed model of how people move in Midtown can be constructed. However, there will be little information on human mobility patterns for the residents in Ridgewood based on only the yellow taxi dataset, which makes the design of such model challenging. A possible approach to deal with the sparse coverage is to fill the gaps using other datasets, such as green taxis (local taxi service in Brooklyn) or data from Uber.

This scenario also shows the need to examine different data slices at different resolutions to help determine the quality of the data: by analyzing the data at the neighborhood level and on an hourly basis, the sparsity issue can be identified, while a coarser resolution may hide this matter. It is thus crucial to have usable tools that enable users to easily and interactively explore these data at multiple granularities.

3.5 Inferring Speed from Trip Duration and Distance

The taxi data can be used as a proxy to understand how traffic flows in New York City. Given the attributes *trip duration* t and *trip distance* d, the average speed associated to each trip can be computed as d/t. Given the average speeds for all trips, along with their spatio-temporal attributes, it is possible to derive, for example, analyses about how traffic jams are distributed in New York City [26]. In this scenario, it is important to discard outliers for trip duration and trip distance as they will negatively affect the computation of average speeds. While entries with incorrect values (negative or zero values) are easy to identify, detecting which positive valued outliers should be removed is challenging.

Figure 7 shows how average speeds are distributed in the 2011 taxi dataset. The speed limit in New York

City was 30 miles per hour in 2011,[4] but there are taxi trips in the dataset that surpassed such limit. Deciding which of them correspond to data inconsistencies, and which simply correspond to drivers traveling over the speed limit, is a difficult task. In Figure 7, while most results look valid, as speeds between 30 and 50 miles per hour probably correspond to real occurrences, values above 100 miles per hour are likely to correspond to errors in the dataset.

Before deciding which trips should be removed, it is necessary to remove trips that are inconsistent, i.e., trips having attributes d or t equal to zero. Poco et al. [26] showed that these trips carry a significant negative impact on speed computations and general traffic flow analysis. After removing these trips, one can address the problem by using a combination of traditional outlier detection techniques and domain knowledge. For outlier detection, it is possible to define a standard distribution that should fit the average speed distribution (e.g., a Gaussian distribution), and remove all trips that are a few standard deviations (say 1 or 2) away from the mean. Domain experts can also help uncover behaviors that can be normal, even if they seem to be outliers. It is possible, for instance, that drivers reach high speeds in certain parts of uptown Manhattan when moving to upstate New York roads. As in other cases, slicing the data into spatial regions and temporal ranges, alongside the aid of a domain expert, can be useful to uncover specific speed patterns in New York City.

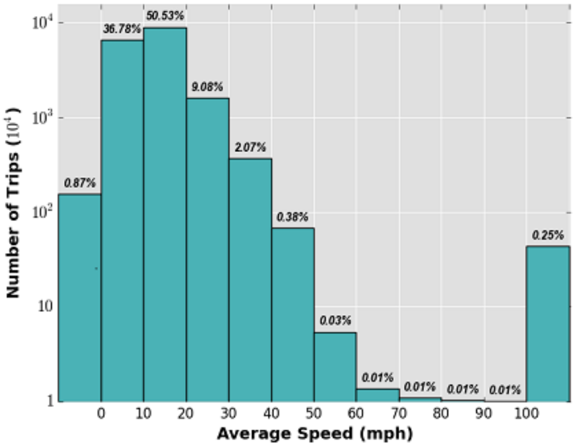


Figure 7: Distribution of taxi average speeds in miles per hour (mph) for the 2011 taxi dataset.

3.6 Inaccurate GPS Readings

GPS readings are not always accurate, especially in cities with a large number of tall buildings. GPS signals are also heavily influenced by the number of GPS satellites: the more satellites are used, the more accurate are the positions. When a taxi passes by a tall building or other obstructions, the set of satellites to which its GPS is associated will likely change. This signal switch between different sets of satellites negatively impacts the position accuracy. The quality of the GPS receiver algorithm for processing the satellite signals might also lead to an inaccurate position.

Figure 8 shows many such errors: taxis in the rivers, in the ocean, and outside North America. Inaccurate GPS points can lead to misleading results. If one wants to detect trendy areas where residents and tourists often go to in NYC, for example by using an algorithm such as k-means, the inaccurate GPS points will lead to meaningless clusters—outside NYC and over the water.

Visualization is an effective mechanism to identify these inconsistencies. By looking at the maps in Figure 8, one can easily see the incorrect locations. To remove GPS inconsistencies, clustering methods can be used. If the geographical boundaries are known in advance, it is possible to check whether they are inside valid polygons. For the NYC taxi data, we can check whether pickups fall within a neighborhood (or zip code) within the city bounds.

3.7 Ghost Trips

While analyzing the taxi data, we discovered a large number of overlapping trips for the same taxi, i.e., for a given taxi, a new trip starts before the previous trip has ended. We call these trips *ghost trips*. The reason behind this data inconsistency is unclear: some trips may overlap due to a device error, or simply because the taxi driver

[4] http://cityroom.blogs.nytimes.com/2011/05/12/a-spooky-reminder-to-obey-the-speed-limit/

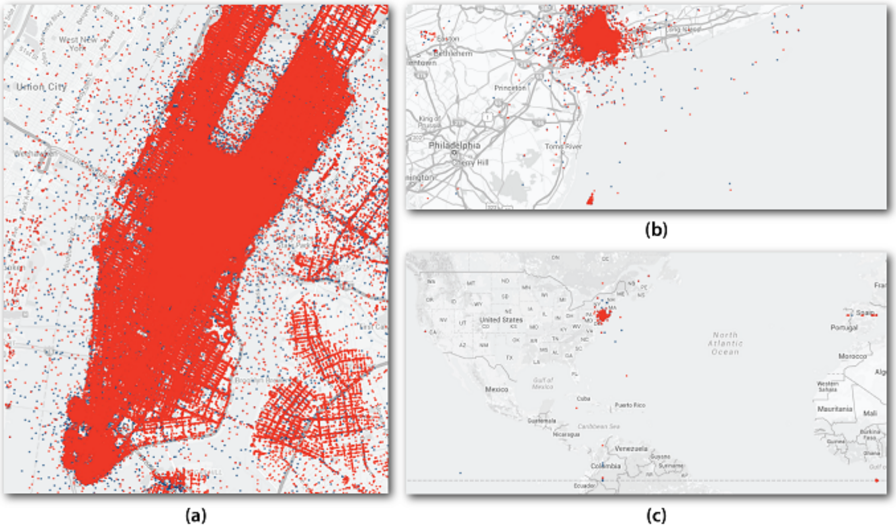


Figure 8: Inaccurate GPS points (a) in rivers, (b) in the ocean, and (c) outside North America.

forgot to log the end of a trip after dropping off passengers. Nevertheless, they certainly affect further analysis on the data, such as data-based human mobility models [42].

In the 2010 taxi dataset, for the month of May, there were 7.1 million ghost trips. Given the 154 million trips that took place that month, this corresponds to an error rate of about 4.60%. To better understand which of the overlapping trips are defective, we would need domain knowledge from expert users and TLC to perform data cleaning: all the trips or just a subset may be erroneous. The number of ghost trips is much smaller for the 2011 dataset: the error rate is only 0.20%. Since the taxi dataset for 2011 has considerably fewer invalid values compared to 2010, as described in Section 2.1, one possible explanation is that different cleaning procedures were used for these two years, and inconsistencies such as ghost trips were removed before the release of the 2011 dataset.

4 Discussion

In this paper, we discussed some of the challenges involved in cleaning spatio-temporal urban data. We presented a series of case studies using the NYC taxi data that illustrate data cleaning challenges and suggested potential methodologies to address these challenges. These methodologies form the basis for integrating cleaning with data exploration. Data cleaning is necessary for data exploration, and through data exploration, users can attain a better understanding of the data which can lead to the discovery of cleaning constraints and enable them to discern between errors and features. Data exploration, however, requires a complex trial-and-error process. Thus, usable tools are needed to guide and assist users in the cleaning process. As the case studies we discussed illustrate, this is particularly true for spatio-temporal data, where visual analytics and event detection techniques at different resolutions are essential to identify quality issues.

The case studies presented in Section 3 show that some cleaning decisions are not clear cut. Often, multiple datasets are required to help an expert decide whether a data point is erroneous or represents an important feature. While there has been preliminary work on the discovery of relationships across datasets [8], there are still many open problems in identifying relevant data that can be used to explain events within a large collection of datasets and in a systematic fashion.

Lack of sufficient knowledge is another issue that hampers data cleaning. Even though experts can (and should) be involved in most of the process, they may be unavailable, or it may be expensive to hire them for cleaning large datasets. Crowdsourcing systems could help the data analyst clean data more efficiently: user feedback can be used to learn features and "separate the wheat from the chaff."

Different questions that arise during exploration may require different cleaning strategies. While visualization helps in identifying potential unusual behavior, other techniques are also necessary in for data cleaning, including automatic event detection and clustering. The fact that these different techniques are applied in trial-and-error fashion underscores the importance of maintaining *provenance* of the cleaning process. Provenance not only enables reproducibility, but it also helps in exploration. As we have shown in previous work, provenance information can be used to support reflective reasoning, to create and refine analysis pipelines by example, to guide the user by providing recommendations for next steps to try, and to perform exploration collaboratively [19, 32, 33]. In addition, provenance provides detailed documentation of all cleaning steps applied to a dataset, and this knowledge is crucial during analyses. For instance, by examining the taxi data, we believe that the years of 2011 and 2012 were better cleaned than, say, the 2010 dataset. This could be confirmed if we had access to the provenance of the cleaning process. In this case, provenance would also allow users to identify which cleaning techniques—applied to newer datasets (e.g., 2012) by the TLC—could be re-used to clean older datasets (e.g., 2010). By applying the same cleaning process to the different datasets, analyses over them would likely be more consistent.

Acknowledgments We thank the New York City TLC for providing the data used in this paper. This work was supported in part by a Google Faculty Award, IBM Faculty Awards, the Moore-Sloan Data Science Environment at NYU, the NYU Tandon School of Engineering, the NYU Center for Urban Science and Progress, and NSF awards CNS-1229185 and CI-EN 1405927. Freire is partially supported by the DARPA MEMEX program.

References

[1] G. Andrienko and N. Andrienko. Spatio-temporal Aggregation for Visual Analysis of Movements. In *Procedings of IEEE Visual Analytics Science and Technology*, pages 51–58, 2008.

[2] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. Visual Analytics Focusing on Spatial Events. In *Visual Analytics of Movement*, pages 209–251. Springer Berlin Heidelberg, 2013.

[3] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. From Movement Tracks through Events to Places: Extracting and Characterizing Significant Places from Mobility Data. In *Proceedings of IEEE Visual Analytics Science and Technology*, pages 161–170. IEEE, 2011.

[4] L. Barbosa, K. Pham, C. Silva, M. Vieira, and J. Freire. Structured Open Urban Data: Understanding the Landscape. *Big Data*, 2(3), 2014.

[5] L. Berti-Equille, T. Dasu, and D. Srivastava. Discovery of complex glitch patterns: A novel approach to quantitative data cleaning. In *Proceedings of the International Conference on Data Engineering*, pages 733–744, 2011.

[6] M. R. Bloomberg and D. Yassky. 2014 Taxicab Fact Book. `http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf`, 2014.

[7] D. Castellani Ribeiro, H. T. Vo, J. Freire, and C. T. Silva. An Urban Data Profiler. In *Proceedings of the International Conference on World Wide Web*, WWW '15 Companion, pages 1389–1394, 2015.

[8] F. Chirigati, H. Doraiswamy, T. Damoulas, and J. Freire. Data polygamy: The many-many relationships among urban spatio-temporal data sets. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, 2016. To appear.

[9] T. Dasu, J. M. Loh, and D. Srivastava. Empirical glitch explanations. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 572–581, 2014.

[10] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. Silva. Using topological analysis to support event-guided exploration in urban data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2634–2643, 2014.

[11] H. Doraiswamy, H. T. Vo, C. T. Silva, and J. Freire. A GPU-based index to support interactive spatio-temporal queries over historical data. In *IEEE International Conference on Data Engineering*, 2016. To appear.

[12] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013.

[13] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science and Engineering*, 10(3):11–21, 2008.

[14] B. Goldstein and L. Dyson. *Beyond Transparency: Open Data and the Future of Civic Innovation*. Code for America Press, San Francisco, USA, 2013.

[15] Y. Gu and C. Wang. itree: Exploring Time-Varying Data Using Indexable Tree. In *IEEE Pacific Visualization Symposium*, pages 137–144, 2013.

[16] M. Hoai and F. De la Torre. Max-Margin Early Event Detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2863–2870, 2012.

[17] J. Höchtl and P. Reichstädter. Linked Open Data - A Means for Public Sector Information Management. In *Electronic Government and the Information Systems Perspective*, volume 6866 of *Lecture Notes in Computer Science*, pages 330–343. Springer, Berlin Heidelberg, 2011.

[18] B. Katz and J. Bradley. *The Metropolitan Revolution: How Cities and Metros Are Fixing Our Broken Politics and Fragile Economy*. Brookings Focus Book. Brookings Institution Press, 2013.

[19] D. Koop, C. E. Scheidegger, S. P. Callahan, J. Freire, and C. T. Silva. Viscomplete: Automating suggestions for visualization pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1691–1698, 2008.

[20] L. Lins, J. T. Klosowski, and C. Scheidegger. Nanocubes for Real-Time Exploration of Spatiotemporal Datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2456–2465, 2013.

[21] E. McFowland III, S. Speakman, and D. B. Neill. Fast Generalized Subset Scan for Anomalous Pattern Detection. *Journal of Machine Learning Research*, 14:1533–1561, 2013.

[22] D. B. Neill and G. F. Cooper. A Multivariate Bayesian Scan Statistic for Early Event Detection and Characterization. *Machine learning*, 79(3):261–282, 2010.

[23] City of Chicago Data Portal. `https://data.cityofchicago.org`.

[24] NYC OpenData. `https://nycopendata.socrata.com`.

[25] M. Ota, H. Vo, C. Silva, and J. Freire. A scalable approach for data-driven taxi ride-sharing simulation. In *IEEE International Conference on Big Data*, pages 888–897. IEEE, 2015.

[26] J. Poco, H. Doraiswamy, H. Vo, J. L. Comba, J. Freire, C. Silva, et al. Exploring traffic dynamics in urban environments using vector-valued functions. *Computer Graphics Forum*, 34(3):161–170, 2015.

[27] W. Rao, K. Zhao, Y. Zhang, P. Hui, and S. Tarkoma. Towards maximizing timely content delivery in delay tolerant networks. *IEEE Transactions on Mobile Computing*, 14(4):755–769, 2015.

[28] R. E. Roth. An Empirically-Derived Taxonomy of Interaction Primitives for Interactive Cartography and Geovisualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2356–2365, 2013.

[29] S. Sarawagi. Explaining Differences in Multidimensional Aggregates. In *Proceedings of the International Conference on Very Large Data Bases*, pages 42–53, 1999.

[30] S. Sarawagi, R. Agrawal, and N. Megiddo. *International Conference on Extending Database Technology*, chapter Discovery-Driven Exploration of OLAP Data Cubes, pages 168–182. Springer Berlin Heidelberg, 1998.

[31] T. H. Savage and H. T. Vo. Yellow cabs as red corpuscles. In *Proceedings of Workshop on Big Data and Smarter Cities*, 2012.

[32] C. E. Scheidegger, H. T. Vo, D. Koop, J. Freire, and C. T. Silva. Querying and creating visualizations by analogy. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1560–1567, 2007.

[33] C. E. Scheidegger, H. T. Vo, D. Koop, J. Freire, and C. T. Silva. Querying and re-using workflows with vistrails. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 1251–1254, 2008.

[34] N. Shadbolt, K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, H. Wendy, and M. Schraefel. Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent Systems*, 27(3):16–24, 2012.

[35] G.-D. Sun, Y.-C. Wu, R.-H. Liang, and S.-X. Liu. A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges. *Journal of Computer Science and Technology*, 28(5):852–867, 2013.

[36] TaxiVis. `https://github.com/ViDA-NYU/TaxiVis`.

[37] TLC Trip Record Data. `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`, 2015.

[38] J. Wakefield and A. Kim. A Bayesian Model for Cluster Detection. *Biostatistics*, 14(4):752–765, 2013.

[39] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. v. d. Wetering. Visual Traffic Jam Analysis Based on Trajectory Data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2159–2168, 2013.

[40] H. Wickham. Tidy Data. *The Journal of Statistical Software*, 59, 2014.

[41] K. Zhao, M. P. Chinnasamy, and S. Tarkoma. Automatic City Region Analysis for Urban Routing. In *IEEE International Conference on Data Mining Workshop*, pages 1136–1142, 2015.

[42] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Nature Scientific Reports*, 5(9136), March 2015.

[43] Y. Zheng, F. Liu, and H. Hsieh. U-Air: When Urban Air Quality Inference Meets Big Data. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1436–1444, 2013.