Profiling for Confidence: Debugging Relationships among Urban Spatio-Temporal Datasets

Laís M. A. Rocha[1], **Mirella M. Moro**[1], **Juliana Freire**[2]

[1]Universidade Federal de Minas Gerais (UMG). Belo Horizonte, MG – Brazil

[2]New York University (NYU). New York City, NY – USA.

{laismota,mirella}@dcc.ufmg.br, juliana.freire@nyu.edu

Abstract. *We aim to help users identify potential issues in* spatio-temporal data *and thus gain trust in the results they derive from such data – a crucial benefit in the era of data science and big data. We propose a framework for profiling spatio-temporal relationships that automatically identifies data slices that deviate from what is expected, which can be further analyzed for quality issues and/or potential effects on analysis' results. We describe the profiling methodology and present cases studies using real urban datasets, then emphasizing the need for spatio-temporal profiling to build trust on data analysis' results.*

1. Introduction

Urban data reflect how urban components behave over space and time, from residents' interactions (e.g., complaints) to infrastructure (e.g., subway and bus lines) and the environment (e.g., weather). These data (usually openly provided by major cities) may enable a better comprehension about the individual components, and provide insights into how they *interact*. Hence, there is growing interest in developing techniques that automatically discover *relationships* among large urban dataset collections (Chapter 2 of this dissertation). But analyzing these relationships and deriving actionable insights are not easy.

A known issue on data trend searching is the Simpson's paradox: a trend that appears in several data groups reverses when these groups are combined [Alin 2010]. In practice, we have observed many instances of a *quasi-paradox*: some spatio-temporal slices of a relationship have a large deviation – compared to others. We call such discrepancies *deviations*. Also, depending on how important these slices are or how often *deviations* occur, a deeper analysis of the relationship and its implications is required. Overall, to explain and trust an analysis result require examining such deviations, as they can provide a nuanced view of an urban process and influence ongoing (future) actions.

Note that deviations may be present in *relationships between datasets*, while no corresponding deviations exist in the value distributions of the *individual datasets*. One possible method to discover deviations in individual data is to compute their outliers, i.e., data points that differ significantly from other observations with respect to the mean of the distribution. While discovering deviations in the individual datasets is an important data analysis task, it is not sufficient for uncovering deviations among relationships.

Challenges on Deviation Discovery. By automatically discovering deviations over space and time, we can help users to focus on *interesting* regions and periods of data relationships. However, due to the inherent complexity in spatio-temporal urban data, identifying

Figure 1. Overview of the profiling framework methodology.

meaningful deviations is challenging: there can be a large number of spatio-temporal slices to examine (due to multiple data resolutions, e.g., or GPS coordinates in seconds); and splitting urban data into meaningful slices requires the data to be aggregated at a spatio-temporal resolution that best fits the data (or that simply makes sense for a certain analysis). However, this may lead to sparsity (i.e., slices with a very small number of data points) and consequently, results that are not significant. In such cases, assessing the statistical meaningfulness of the uncovered relationships is of crucial importance.

Main Contributions. We propose a *relationship-based profiling* methodology to help data scientists understand data as well as results derived from data analysis and their trustworthiness. By profiling the data, we discover potential problems as well as identify valid, useful, and understandable patterns through relationships across urban datasets. Our main contributions are summarized as follows.

1) Relationship-Based Profiling Framework: it automatically detects and assesses relationships associated to individual data slices that significantly deviate from the behavior expected from other slices. These deviations, in turn, can uncover interesting features or issues with the data. It also evaluates the statistical significance of the relationship for all of the derived slices to avoid potentially spurious conclusions from the data.

2) Case Studies: We show its applicability over open urban datasets, and present case studies that clarify problems and challenges. These studies also indicate the relevance of detecting deviations as a means to uncovering potential data quality issues.

3) Debug-Data application: A web application to visualize its profiling results.

Results of this work are published at a major Big Data venue [Rocha et al. 2019].

2. Research Contributions

Next, we *briefly* describe methodology, analyzed datasets and main case studies. More information is at the dissertation text in informed chapters and sections in parentheses.

2.1. Spatio-Temporal Profiling Framework (Dissertation Chapter 3)

Figure 1 shows a high-level overview of the proposed profiling framework. Given a collection of datasets, the framework starts by aggregating them at different temporal and spatial resolutions. Then attribute values are normalized (avoiding metric distortions). Relationships are computed at different levels of granularity with respect to both time and space. The statistical significance of the derived relationships is assessed. Finally, it identifies deviations and uses different measures to evaluate them.

Data Aggregation. Urban datasets come in different spatio-temporal resolutions (e.g., Table 4.1 in the full text). To compute relationships across them, the framework first aggregates each attribute into a set of default resolutions, which can be defined to match

an application's requirements (Section 3.1.1). Aggregating uses different functions depending on attribute type and on the analysis goal. Then, the framework adds a *density attribute* to keep track of the number of data records for each location in time and space. This value is used to evaluate the deviations identified in relationships (Section 3.4).

Data Normalization. Attribute values come in different units and scales, which may cause metric distortions when computing relationships. Hence, the framework normalizes attributes using *z-scores*, also known as *standard scores*. This method preserves ranges (maximum and minimum) and accurately represents the dispersion of values of an attribute around their mean (Section 3.1.2).

Relationship Computation. Data are imperfect by nature, and their analyses may result in erroneous conclusions, which in turn may lead to bad decisions. Intuitively, if a pair of datasets has many slices in which a given relationship behaves differently when compared to the whole data, or some areas show a large deviation from the expected distribution, there is evidence that the user must further investigate before drawing broad conclusions. We use three correlation metrics to identify a possible relationship between a pair of normalized attributes: Pearson Correlation (PC), Spearman's Correlation (SC), and Weighted Linear Regression (*WLR*) (Section 3.2). With spatio-temporal data, density may be a problem in spatial resolutions where data is too sparse. To model the importance of values based on their density, we propose two novel weighting functions for *WLR*: *LR_HDHW* and *LR_CPP* (Section 3.2.1), which emphasize points with higher density, i.e., to points associated to more data records.

Result Grouping. The used correlation metrics are agnostic to spatio-temporal resolutions; hence, to capture data variations when computing correlations, we group the detected relationships by space (*space perspective*) and by time (*time perspective*). For each perspective, we compute correlations in a *local* level and in a *global* level, to be compared later. For instance, on a space perspective, the global correlation is computed over the entire spatial resolution (e.g., city), and the local ones are computed over finer-grained regions (e.g., neighborhoods). On a time perspective, the global correlation is computed over the entire temporal resolution (e.g., all years for which the data is available), and the local ones are also computed over finer-grained temporal ranges (e.g., each year).

Statistical Significance. To get the statistical significance of the generated relationships, we compute the T-test [Kalpić et al. 2011] for all the correlation functions supported (PC, SC, and WLR). Given the framework evaluates thousands of relationships, we need to address the problem of *multiple comparisons* for a more reliable profiling, and to prune potentially spurious results. Hence, we use the BH procedure [Benjamini and Hochberg 1995] as it is a standard controlling procedure known to have greater power, i.e., less probability of producing false negatives (Section 3.3).

Deviation Evaluation. Given two input datasets, the framework proceeds as aforementioned. The last step evaluates deviations in relationships that may be linked to different factors (e.g., many slices with correlations harshly deviating from the expected for that relationship; or areas with local correlations deviating much from the global one). We use the *standard difference* metric to assess deviation harshness. After calculating *std-diff* for different perspectives, we sort the values in descending order. The resulting rank highlights the most deviating regions, which require attention from analysts (Section 3.4).

Meaningful Deviations. Detecting *meaningful deviations* among relationships is a difficult task. Statistical significance tests and pruning using the BH procedure are some of the main tools used to filter out relationships that *are not likely* to be relevant. Besides statistical significance, we address the density problem. Specially in urban data, we find many low-density spatial and/or temporal regions as we aggregate and split the data before computing the relationships. We filter regions without enough density by calculating the average of the density values for different spatio-temporal resolutions (Section 3.4.2).

2.2. Case Studies: Inspecting Urban Data (Dissertation Chapter 4)

The spatio-temporal nature of urban data enables analyzing relationships in multiple slices. Data can be aggregated into different spatial (e.g., neighborhoods, zip codes) and temporal resolutions (e.g., hourly, monthly). Depending on the resolution, interesting data slices may be easily identifiable or completely hidden. Also, spatio-temporal patterns present other challenges: data considered unreliable in a specific time period or spatial region (due to large deviation) may in fact be an important feature or event.

We use our framework over various case studies that expose the challenges involved in identifying potential quality issues when analyzing spatio-temporal correlations. We use seven urban datasets from NYC, mostly obtained from the NYC Open Data portal and different NYC Agencies. Each dataset consists of a metadata and a set of spatial, temporal, numerical, and identifier attributes (Section 4.1). We group case studies in: those with unusual local correlations filtered out for not being statistically significant; those with uncertain reason for deviating local correlations; and those whose reasons for deviations may be outside the data and not be noticeable if the corresponding attributes are analyzed individually, as summarized next.

The Importance of Statistical Significance. Many relationships between different attributes from the 311, Weather and Vehicle Collisions datasets are not statistically significant. Hence, while some of these relationships have a moderate or strong correlation, they are likely random or coincidental. This emphasizes the importance of considering statistical significance when evaluating relationships across spatio-temporal data. (Section 4.2).

Deviations can be a Data Quality Issue or a Feature. When analyzing many relationships across different urban datasets and looking for trends in data, Simpson's paradox may occur. Another version of Simpson's Paradox is when a correlation in one direction in stratified groups changes direction in aggregated groups. We argue such reversal may not occur completely or that the directions of relationships may not be opposite. However, in some cases, the deviations are large enough to raise questions about a certain previously established relationship or trend.

The neighborhood resolution has an interesting relationship between *Weather Temperature* and *Airbnb Daily Prices* when using *space* as grouping perspective. Positive correlations were detected with all correlation metrics (global correlation around 0.50 from 2015 to 2018), suggesting a possible increase in prices during seasons that attract more tourists, i.e., when the weather is more pleasant. However, all of the correlation metrics found the same Top-3 outlying neighborhoods with negative relationships: *Sunset Park*, *Flatbush*, and *Fort Greene*. These neighborhoods present deviations w.r.t. the city (global spatial resolution) in 2015. For instance, Sunset Park presents a reversal with a local correlation of -0.52, and almost five standard deviations above the mean. Flatbush

has a complete reversal with a local correlation of -0.3. Figure 4.5 (Dissertation) shows a comparison between the local correlation of each neighborhood and the global one for the city over the years: both have a higher deviation in 2015 that decreases until 2017.

The question then becomes if the deviation is due to a data quality issue, or if it is a consequence of an odd event that happened on that particular location and time. As the temperature dataset is uniform for the whole city, i.e., it cannot be spatially sliced for different boroughs or neighborhoods, we reckon that these unexpected correlations are related to the Airbnb dataset. Regardless of the perspective used to group the results, assessing the cause of deviations is hard. Common ones are: missing data, and a characteristic or an event occurring in a certain space and time.

Nonetheless, such unusual deviations only become apparent for the relationships between the datasets, i.e., *they do not correspond to harsh deviations in the Airbnb dataset*. We searched for outliers in the data distribution of the Airbnb Price dataset in 2015, computing the number of standard deviations above/below the mean for all points. Then, other neighborhoods, different from those found with our framework, are identified as outliers. This case reinforces the importance of not only analyzing the data individually, but also *using a profiling method based on relationships*. Such deviations in relationships, whether reversed or not, may be useful for highlighting possible problems in the data, features or interesting events about a given relationship (Section 4.3).

Features May Be Outside the Data. Analyzing relationships between datasets may uncover positive or negative trends. However, if the data sample does not cover the context in which these relationships unfold, a user may draw wrong conclusions. In these situations, there could be features outside the analyzed datasets that are crucial for a better understanding on how they relate. A more complete analysis may then need integrating such datasets with others; but even this step turns out to be challenging. The interaction between Crime and Yellow Taxi datasets illustrates such a case (Section 4.4).

General Insights. We demonstrate the usefulness of our framework through a series of case studies motivated by urban data and important elements that impact the operation of a city. Regions designated as having the highest deviations depend on each correlation metric used. The method evaluates deviations between local and global correlations and uses such deviations as a parameter to indicate possible anomalies or characteristics in relationships. Hence, the choice of the correlation metric can directly interfere in the profiling and the deviations found. These studies show how our relationship-based profiling enables domain experts to analyze data to better understand them and ensure reliability for future analyses. They also show how this profiling can help users identify and reason about significant deviations, possible quality issues and features in spatio-temporal data and possible features outside the data. In addition, they present challenges involved in identifying potential quality issues in such complex and large data. Finally, they demonstrate the importance of detecting deviating relationships as a means of uncovering potential data quality issues to provide better confidence to the user.

2.3. Web Application (Dissertation Chapter 5)

The framework has an application to test and visualize its profiling results. It shows the top-k *deviations* for an analytical tasks and the results based on different correlation metrics, for each relationship. It allows to visualize the spatial and temporal aspects of

the results to help users finding possible datasets (or slices) that are potentially unreliable, easily.

3. Conclusion

Here, we combined data profiling and relationship analysis to propose a relationship-based profiling framework. Unlike previous work, our method profiles several spatio-temporal datasets and uses their relationships to detect significant deviations between local and global correlations. The framework can be used by experts to better understand any type of urban interactions and obtain higher confidence in the analyzed data. The main benefit is that these deviations may help uncover potential anomalies or features in the data through relationships. More importantly, this dissertation emphasizes serious questions on available urban data, which could impair any analysis over those and jeopardize any decision taken based on such analyses. Also, it takes advantage of Computer Science techniques to better find data issues not easily detected any other way than through analyzing relationships from distinct datasets, which our framework efficiently does. The contributions of this study also surpass the Computing field, as it points out data issues that must be handled by Urban planners, city administrators, and other professionals. Overall, Data Scientists and Engineers clearly need to review their process of collecting and processing data from multiple sources, before making them available and prone to errors. This dissertation is part of a major project (see Acknowledgements), involving urban data, discovering relationships between spatio-temporal datasets and outliers detection. This study considerably increased the reliability and quality of the data used in the project research, mostly affecting two ongoing PhD Theses, and allowed a more accurate and comprehensible analysis and future diagnosis of the results. For future work, we plan to explore new correlation metrics and machine learning prediction algorithms and analyze correlations and deviations over more than two attributes at a time, and improve our deviation evaluation.

Acknowledgements. Work partially supported by CNPq and FAPEMIG (Brazil), the U.S. National Science Foundation under grant OAC-1640864, the DARPA D3M program, the Gordon and Betty Moore Foundation, and the Sloan Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of funding agencies.

References

Alin, A. (2010). Simpson's paradox. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(2):247–250.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Kalpić, D., Hlupić, N., and Lovrić, M. (2011). Student's t-tests. *International encyclopedia of statistical science*, pages 1559–1563.

Rocha, L. M., Bessa, A., Chirigati, F., OFriel, E., Moro, M. M., and Freire, J. (2019). Understanding spatio-temporal urban processes. In *IEEE Big Data*, pages 563–572.