Human-in-the-Loop Integration of Complex and Noisy Data (VAST MC2)

Daniël M. Bot[¶,*]
I-BioStat, Data Science
Institute, Hasselt University,
Hasselt, Belgium

Jannes Peeters[¶,†]
I-BioStat, Data Science
Institute, Hasselt University,
Hasselt, Belgium

Danai Kafetzaki[‡]
I-BioStat, Data Science
Institute, Hasselt University,
Hasselt, Belgium;
Sagalassos Archaeological
Research Project, KU
Leuven, Leuven, Belgium

Jan Aerts[§]
I-BioStat, Data Science
Institute, Hasselt University,
Hasselt, Belgium

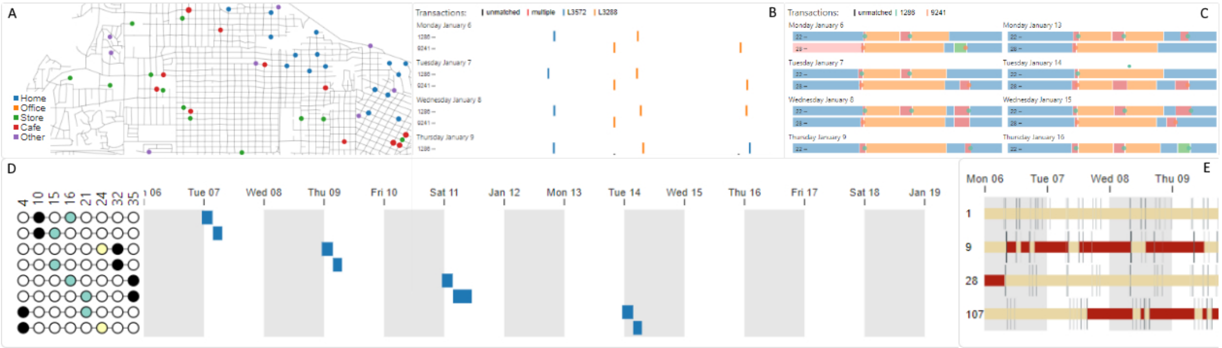


Figure 1: Screenshots of our interfaces: **A** Annotated *points of interest* (POIs) coloured by type. **B** Transactions of credit cards over time coloured by matching loyalty card. **C** Stationary periods of cars coloured by POI type with transactions coloured by credit card. **D** UpSet visualisation [2] indicating cars present at the same POI with an interactive timeline. **E** Timelines of GPS activity indicating periods in which cars were stationary (yellow), moving (black), or data was missing (red).

Abstract

When considering complex and noisy data, fully automated data integration is not straightforward. In this paper, we illustrate the role of visual analytics in this process by presenting our submission to the yearly VAST challenge. We show how the human-in-the-loop approach helps distinguish patterns due to systematic or specific errors in the data from patterns that represent actual behaviours and phenomena.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics

1 Introduction

In the VAST Challenge 2021's second mini-challenge[1], contestants were asked to identify clues regarding the disappearance of individuals from a fictional company. Specifically, the data consisted of geospatial tracking data of company cars, a car assignment table, a touristic map of the island that the company is located on, as well as credit card and loyalty card transactions of the two weeks leading up to the employees' disappearance. The goal was to identify suspicious behaviours or patterns.

In this paper, we present our methodology for solving the mini-challenge. Throughout the project, we developed visual tools[2] to understand the data and its anomalies and guide our data processing, integration and analysis decisions. We preferred human-in-the-loop visual analytics approaches over fully computational solutions because we found them easier to develop, understand, and trust for the tasks in this challenge.

2 Methods and Findings

Three of the provided data sources share time as a common attribute. However, their *granularity* differed: GPS traces were accurate to the second, credit card transactions to the minute, and loyalty card transactions to the day. In the terminology of Aigner et al. [1], we mainly focused on the *states* over time rather than the *events* which indicate state changes. For example, the GPS data indicates when cars were moving (an *event*), which changes the car's position (the *state*).

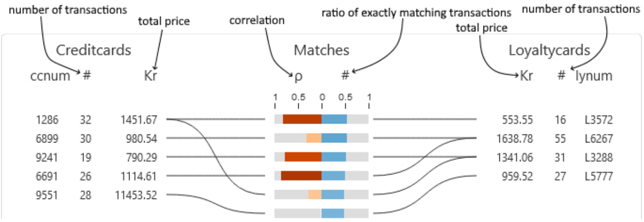
Several relations were present in the data. The transactions indicate the time at which cards were used at particular businesses. In addition, a car-assignment table was provided, indicating which employee used which car. Other relations had to be uncovered. First, we matched the transactions of loyalty cards and credit cards, assuming each employee has one of each but allowing for more complex relations. Then, we simultaneously matched cars to loyalty-credit card matches and businesses to GPS positions where cars were stationary—i.e. *points of interest* (POIs). Finally, we analysed meetings of people, looking for suspicious patterns.

2.1 Credit-to-loyalty-card matching

We used two metrics for this matching process: (1) the correlation between vectors indicating the total amount of money spent at each business on each day, and (2) the Jaccard index of card's transaction sets, where transactions are equal when they occur at the same business on the same day for the same price. Matches were visualised as a bipartite graph, with coloured bars indicating the metric values (Fig. 2) beside a detail-view (Fig. 1.B) showing each selected credit card's transactions on a time-axis coloured by the matching loyalty card in a small multiple for each day. The detail view indicated credit

[*] e-mail: jelmer.bot@uhasselt.be
[†] e-mail: jannes.peeters@uhasselt.be
[‡] e-mail: danai.kafetzaki@uhasselt.be
[§] e-mail: jan.aerts@uhasselt.be
[¶] These authors contributed equally to this work.

[1] https://vast-challenge.github.io/2021/MC2.html
[2] https://vda-lab.github.io/2021/07/vast2021



Creditcards		
ccnum	#	Kr
1286	32	1451.67
6899	30	980.54
9241	19	790.29
6691	26	1114.61
9551	28	11453.52

Loyaltycards		
Kr	#	lynum
553.55	16	L3572
1638.78	55	L6267
1341.06	31	L3288
959.52	27	L5777

Figure 2: A partial screenshot of the credit-to-loyalty-card matching graph.

card transactions with multiple matches in red and connected credit card transactions that matched the same loyalty card transaction, allowing us to detect and solve these conflicts.

Using our interface, we discovered two data issues: transactions for one particular business always occurred one day earlier in the loyalty card data than in the credit card data, and some credit card transactions were precisely 20, 24, 60, or 80 units higher than their only potential matching loyalty card transaction.

2.2 GPS pre-processing

The GPS data contained traces of car positions. The GPS trackers were only active when cars moved, so we looked at the periods between samples to identify stationary states. We employed a simple decision tree to classify these time periods as moving (duration $<$ 30 s and average speed $>$ 20 km/h), or stationary (distance travelled $<$ 220 m), or missing (otherwise).

An interface containing a street map and GPS timeline (Fig. 1.E) was used to visualise all GPS samples enclosing non-moving periods, a red line connected the samples around missing periods. We found car 28's measurements were noisy and had a fixed translation error. Transaction times could potentially be used to fill in some of the missing periods manually.

The same interface was used to manually mark *points of interest* (POIs) (Fig. 1.A) , indicating all distinct places where cars were stationary within the two weeks. POIs were annotated with a type based on the tourist map and the average number of cars present throughout a day. Especially homes are easy to identify, having a U-shaped *busyness trace*.

2.3 Car to credit card matching

Two relations remain unknown: (1) which POI belongs to which business? and (2) which cards are used with which car? These relations had to be uncovered simultaneously, as they constrained each other's possible matches. We used two metrics: (1) the correlation of binary vectors indicating whether a car was stationary or a credit card was used for a transaction in 5-minute intervals, and (2) the precision and recall of a car's stationary periods as a predictor for a credit card's transactions, where transactions had to fall within the stationary period and the stationary period's POI had to match the business (when known). Both metrics excluded stationary periods at homes and the office.

An interface showing the matches as in Fig. 2, a street map with POIs (Fig. 1.A), and a detail view visualising selected matches (Fig. 1.C) was used for the matching process. The detail view shows stationary periods as rectangles coloured by POI type and transactions as circles coloured by credit card, positioned on a time-axis using small multiples per day. Transactions of high-rated matches were manually assigned to stationary periods, introducing POI-to-business constraints and removing the transaction and stationary period from consideration in other matches.

During the matching process, we came across several issues with the transaction times. The credit card transaction times at one particular business were roughly 12 hours too late. In addition, transactions at several breakfast places were all recorded at 12:00 while cars were (mostly) at the office. We discovered that cars with these transactions often stopped at the same POI on their way to work in the morning. Therefore, we used their departure times at these POIs to estimate the transaction times. The same idea was applied to non-credit-card transactions that only had a date when there was a single potentially matching stationary period. Finally, we used the interface to annotate the homes of all cars based on the POI they stayed most nights, discovering households where employees cohabit and some suspicious behaviour.

2.4 Suspicious patterns

We used an UpSet visualisation [2] (Fig. 1.D) with an interactive meeting timeline to analyse meetings between employees. The UpSet visualisation contains circles in a grid; each column corresponds to a car, and each row corresponds to a unique combination of cars with at least one meeting at a POI. Circles are filled in (coloured by household for multi-employee households, black for single-employee households) when a car was present in that row. The timeline shows each period where only the row's cars were at the same location as rectangles coloured by POI type. Tooltips provided more detail: the names of people present, the events (arrival or departure) that mark the start and end of the meeting, and a miniature visualisation of the individual car's stationary periods (with matched transactions). This latter miniature visualisation allowed us to estimate whether people met at a place or just ran into each other. In addition, the interface contained the POI-map (Fig. 1.A) and several options to filter and sort based on POI, car, household, total meeting duration, and meeting onset.

Using this interface, we found a surprise party for a particular employee on Friday the 10th, some kind of nightly guard duty at executives' homes (Fig. 1.D), two employees who meet for long lunches at the hotel, and executives who played golf together. In addition, we found suspicious truck activity on Thursday the 16th when they drive around in loops without stopping and several more expensive than expected transactions.

3 Conclusion

We presented our approach to VAST challenge 2021's second mini-challenge. Our work heavily relied on visualisations to understand the data, judge our decisions, and discover data problems and their solutions. We used interactive visualisations to manually solve problems that are difficult to automate given the complexity and amount of data provided in the challenge. We believe our human-in-the-loop approach succeeds in providing quick insights, and our interactive visualisations make it easy to design the analyses needed to gather formal support for our findings.

Acknowledgments

This work was supported in part by Hasselt University BOF grant ADMIRE [BOF21GP17], Flemish Government programme "Onderzoeksprogramma Artificiele Intelligentie (AI)", the Research Fund of the University of Leuven [C14/17/025] and the Research Foundation Flanders [G088319N].

References

[1] W. Aigner, A. Bertone, S. Miksch, C. Tominski, and H. Schumann. Towards a conceptual framework for visual analytics of time and time-oriented data. In *2007 Winter Simulation Conference*, pp. 721–729, 2007. doi: 10.1109/WSC.2007.4419666

[2] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, 2014. doi: 10.1109/TVCG.2014.2346248