Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org
IMDEA Networks Institute
Leganés, Spain
Univ. Carlos III de Madrid
Leganés, Spain

Marius Paraschiv
marius.paraschiv@imdea.org
IMDEA Networks Institute
Leganés, Spain

Nikolaos Laoutaris
nikolaos.laoutaris@imdea.org
IMDEA Networks Institute
Leganés, Spain

ABSTRACT

Spatio-temporal information is used for driving a plethora of intelligent transportation, smart-city and crowd-sensing applications. Data is now a valuable production factor and data marketplaces have appeared to help individuals and enterprises bring it to market and the ever-growing demand. Such marketplaces are able to combine data from different sources to meet the requirements of different applications. In this paper we study the problem of estimating the relative value of spatio-temporal datasets combined in marketplaces for predicting transportation demand and travel time in metropolitan areas. Using large datasets of taxi rides from Chicago, Porto and New York we show that simplistic but popular approaches for estimating the relative value of data, such as splitting it equally among the data sources, more complex ones based on volume or the "leave-one-out" heuristic, are inaccurate. Instead, more complex notions of value from economics and game-theory, such as the Shapley value, need to be employed if one wishes to capture the complex effects of mixing different datasets on the accuracy of forecasting algorithms. This does not seem to be a coincidental observation related to a particular use case but rather a general trend across different use cases with different objective functions.

CCS CONCEPTS

• **Information systems → Geographic information systems**; • **Human-centered computing → Ubiquitous and mobile computing**.

KEYWORDS

Shapley Value, Value of Information, Spatio-Temporal Data, Data Marketplaces

ACM Reference Format:
Santiago Andrés Azcoitia, Marius Paraschiv, and Nikolaos Laoutaris. 2022. Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. In *The 30th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '22), November 1–4, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3557915.3561470

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGSPATIAL '22, November 1–4, 2022, Seattle, WA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9529-8/22/11.
https://doi.org/10.1145/3557915.3561470

1 INTRODUCTION

Data-driven decision making is bringing significant improvements to many sectors of the economy, including in several applications related to ubiquitous computing in the areas of transportation, mobility, and crowd-sensing. A solid body of research has studied matters of route optimization and city infrastructure planning [5, 14, 38], whereas companies are increasingly deploying and operating sophisticated systems for optimising their operations using live data. Such models and algorithms often require combining data from different sources and domains.

Data is by now considered a key production factor, comparable in importance to labour, capital, and infrastructure. Companies often need data from third parties, and for this they resort to data marketplaces. There are different types of marketplaces [8, 33, 34]. Personal Information Management Systems (aka PIMS, like Digi.me, Swash, or Meeco) allow individuals to sell their personal data, including their location, whereas general-purpose (AWS) and domain-specific marketplaces for spatio-temporal data allow companies to sell data to other companies, either integrated in already-existing services (HERE, CARTO, ESRI), as real-time streams (Streamr, IOTA, GeoDB), directly downloading datasets, or allowing access to them.

In almost all commercial marketplaces, pricing is left to sellers and buyers to agree. Sellers may set a fixed price, or let buyers bid for data [23], or even do a combination of the two. Such empirical pricing operates with minimal information, namely a high-level description of the dataset, including the number of data points it contains. The research community has already proposed different marketplace architectures to deal with AI/ML tasks, and industry-led initiatives aim to design trustworthy data spaces to share data [9, 17]. In a nutshell, the marketplace is able to train a model [2, 13] or to run code [1, 29] from potential buyers on data provided by sellers. It also ensures that data is accessed according to the terms agreed by both parties, that no data is leaked or replicated, that the intellectual property of buyers is protected, and that transactions and data usage are tracked and accounted.

Contributions: Our work looks at this open problem for the case of spatio-temporal data. In particular, we study how to compute the relative value of different spatio-temporal datasets used in i) forecasting future demand for a service across space and time, and ii) forecasting the travel time between two points A and B in a metropolitan area. Companies already offering service in overlapping areas can, for example, pool together their data to increase the accuracy of forecasting and its coverage. Improved forecasting can be used by the same companies to improve operations, such as dispatching vehicles, providing consumers with better information,

or provisioning service points. It can also be sold to third parties often after bidding and bargaining to agree on the price. In the latter case, the relative value that we compute for each contributing source provides a fair way for buyers to select the most suitable data sources and for marketplaces to split revenue among them.

For the purpose of our work, we concentrate on vehicle-for-hire demand prediction in Chicago and New York, and car travel time forecasting in Porto. While our examples and findings are specific to these particular urban mobility use cases, the methods that we apply for assigning value to spatio-temporal data are more general in scope, and can thus be used in other use cases beyond transportation, such as tourism, health services, entertainment, energy or telecommunications. We answer questions such as "Does combining multiple datasets of past taxi rides always benefit the forecasting accuracy of future services?". Also, when it does, "How should we attribute the improved forecasting precision to the individual datasets used to produce it?".

To do so, we use the Shapley value [32] from collaborative game theory as a baseline metric for establishing the importance of each *player* (be they taxi companies or individual drivers) in the context of a *coalition* of data providers. The Shapley value has many salient fairness properties and wide market adoption, but at the same time entails serious combinatorial complexity challenges since its direct computation in a coalition of size N requires enumerating and calculating the value of $O(2^N)$ sub-coalitions. This may be possible for a few tens of data providers, which is the case of companies in wholesale markets, but becomes impossible when considering hundreds or thousands of them in a retail data market setting.

Furthermore, we look at the trade-off between fairness and scalability/practicality by studying and comparing against simpler heuristics used to estimate the value of data, based on:

- *data volume*, in our case taxi rides, has been used in marketplaces trading marketing or user profiling data [23]. While certainly more practical, this assumes that every ride has equal value for predicting demand or travel time.
- *leave-one-out (LOO).* LOO has been used for "denoising" datasets, by omitting data points that reduce the accuracy of a ML algorithm [18]. Unlike the Shapley value, LOO is examining only a single sub-coalition per source.
- measures of the coverage and the amount of information carried by a dataset such as Shannon's *entropy* [25].
- *similarity metrics* that compare inputs to the aggregate dataset of the whole coalition, often used in detection of outliers [19].

Findings: We first study the value of data fusion at the granularity of companies. Since the number of such companies covering the same geographical area is typically small, the relative value of their data can be computed directly from the definition of the Shapley value. This, however, becomes infeasible at the level of individual taxi drivers, since the latter may amount to several thousands for large metropolitan areas. To address this issue, we compare different approximation techniques, and conclude that an *ad hoc* version of structured sampling [15] performs much better than other approaches such as Monte Carlo [18] and random sampling.

By applying our model and valuation algorithms to taxi-ride data from Chicago, Porto and New York, we find that sufficiently large companies hold enough information to independently predict the overall demand, at city level, or in large districts, with over 96% accuracy. This effectively means that inter-company collaboration does not make much sense in such cases. On the contrary, companies have to combine their data in order to achieve a sufficient forecasting accuracy in smaller districts. We compute the relative value of different contributions in such cases by computing the Shapley value for each taxi company. We find that the values differ by several orders of magnitude, and that the importance of the data of a given company can vary as much as $\times 10$ across districts. More interestingly, the Shapley value of a company's dataset does not correlate with its volume, i.e., some companies that report relatively few rides have a larger impact on the forecasting accuracy than companies that report many more rides. The LOO heuristic also fails to approximate the per company value as given by Shapley.

Similar phenomena are observed at the finer level of individual drivers. We show that by combining data from relatively few drivers one can easily detect peak hours at city level. At district level, however, more data needs to be combined, and this requires making use of our fastest approximations of the Shapley value based on structured sampling. Moreover, using trajectory data from taxis in Porto, we observe again, this time for estimating the travel time within a city, that the value of information contributed by each driver may vary wildly, and that it cannot be approximated based on the volume of rides they report nor via the LOO heuristic.

Overall, using multiple datasets, different forecasting objectives, and at different granularities, our work shows that computing, even approximately, the Shapley value seems to be a "necessary evil" if one wants to split fairly the value of a combined spatio-temporal dataset. Simple heuristics based on volume and LOO fail to approximate the results produced via the Shapley value. Other heuristics tailored to each problem, such as the similarity to the aggregate when predicting demand, or spatio-temporal Shannon entropies when predicting travel time in a city, seem to be doing a better job at approximating Shapley.

2 METHODOLOGY

2.1 Definitions and problem statement

Let N denote a set of data sources, each one contributing a dataset $S_n, n \in N$. A dataset is a set of spatio-temporal observations (x, t) denoting the spatial (x) and temporal (t) coordinates over a common period and geographical area describing the trajectory of taxi rides.

Such dataset is then split into training and test sets for experimentation purposes. Throughout the paper, we train predictive algorithms $(\mathcal{M})$ on subsets of the complete training set (containing a part of the total number of data sources) and perform predictions on the test set. The accuracy of the trained model is gauged by similarity metrics that define the notion of *value* of a dataset S_K, where $K \subseteq N$, which we denote as $v(S_K)$. Thus, $v(S_K)$ represents the accuracy of the predictive model, according to the chosen metric, when training is performed on the data from all sources $k \in K$, and prediction is performed on the fixed test set, containing data from all sources. In a real setting, data buyers would provide the model, the accuracy metric and the test set, whereas the marketplace would help find a combination of suitable data from their sellers to improve the model accuracy.

Our objective is to find a value assignment method, $\mu(n_i)$, that captures the relative importance of the data originating from source

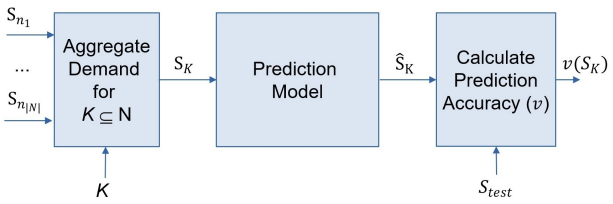


Figure 1: General prediction model.

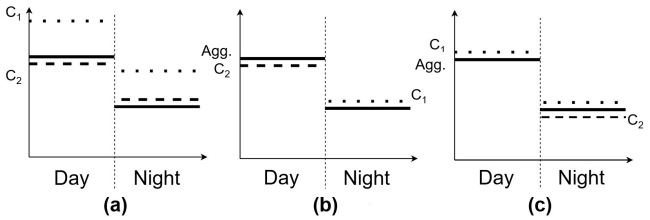


Figure 2: Data aggregation can influence its value in nontrivial ways.

n_i to the predictions of the forecasting algorithm. The value assignment method $\mu(n_i)$ would thus be a function of $\mathcal{M}$, the dataset of source n_i, the datasets of all other sources other than n_i, and v.

$$\mu(n_i) = f(S_{n_i}, \{S_j\}, \mathcal{M}, v), j \in N - \{n_i\} \quad (1)$$

Knowing $\mu(n_i)$ in advance is also useful for buyers to select the most suitable sources, i.e., those with higher $\mu(n_i)$ for their task [7].

2.2 Spatio-temporal forecasting

Figure 1 shows a block diagram that describes the general prediction model used throughout the paper. The model constructs the aggregate input (or training) demand S_K from a set of sources $K \subseteq N$. This input then drives a prediction model that produces a forecast $\hat{S}_K$ which is compared to the ground truth S_N using the similarity metric v. In our definition, the notion of *data value* is directly linked to the test data (S_N) and the similarity metric (v) provided or selected by the buyer in order to compute the model accuracy. The model can easily accommodate any similarity measure as the value function. Throughout the paper, we will use cosine similarity and R^2 score as similarity metrics for predicting travel time and transportation demand, respectively. Moreover, we have validated the obtained outcomes against two other metrics (numerical similarity and *Dynamic Time Warp*), for which we have obtained agreeing results (R^2 of 0.92 for 5 different scenarios).

2.3 Introducing the Shapley value

Establishing individual player contributions to a collaborative game has long been a central problem of cooperative game theory. To this end, Shapley proposed that a player's value should be proportional to their average marginal contribution to any coalition they may join [32]. Even though other notions of *fairness* have been proposed, such as the core[39], we focus on the Shapley value, since it is by far the most widely used by the research community. In what follows, we adapt its definition to our use case.

Let N be a set of sources and S_N be their aggregate data, with a value $v(S_N)$. The *Shapley value* is a uniquely determined vector of the form $(\phi(n_1), ..., \phi(n_{|N|}))$, with $n_1, ..., n_{|N|} \in N$, where

$$\phi(n_i) = \sum_{K \subseteq N \setminus \{n_i\}} \frac{|K|!(|N| - |K| - 1)!}{|N|!} [v(S_K \cup S_{n_i}) - v(S_K)], \quad (2)$$

and $K \subseteq N \setminus \{n_i\}$ takes the value of all possible coalitions of sources, excluding n_i. $v(S_K \cup S_{n_i})$ represents the value of the combined data from the K sources and source n_i. We can use the Shapley values, according to Eq. 1, as a credit assignment method, where $\mu(n_i) = \phi(n_i)$.

2.3.1 A toy example. Consider a group of taxi companies agreeing to pool together their spatio-temporal data, containing demand for taxi rides within a city. One method to determine the value of a company is to observe how well the company is able to reconstruct the total aggregate, that is, the data coming from all companies, by solely using its own. As such, the data of one single company, or a group thereof, is used to train a predictive model, and the reconstruction error, between the prediction and an actual ground truth, is measured. This error, or rather its opposite, the reconstruction accuracy, represents the value of that company or group.

Aggregation leads to a highly non-trivial behavior of the value function, which we illustrate with the toy example depicted in Fig. 2. A number of companies combine their data, to produce a spatio-temporal output or signal (continuous line), representing the total aggregate demand. For simplicity, the time scale is that of a single day, split into day-time and night-time, and also all signals are drawn as constant. Companies whose overall behavior is closer to the *average* may be able to predict the complete aggregate signal by themselves, without a need to form coalitions with other companies. As such, their value will be ranked high by our algorithm. In the example of Fig. 2(a), company C_1 is less valuable than C_2, as the signal of C_2 better emulates the total aggregate.

In the same setting, we also discuss the problem of complementarity, depicted in Fig. 2(b). Company C_1 is only offering its transport services during the night, while company C_2 is active solely during the day. Taken individually, the data of neither of these two is able to reconstruct the complete aggregate. However, they gain tremendous value as a coalition, since the resulting signal covers the entire time-span of the aggregate.

Data aggregation, however, does not always lead to an increase in value. A simple example is presented in Fig. 2(c), one company, C_1 provides data spanning the entire day (both day-time and night-time), and is also close to the total aggregate, while the other, C_2, only provides data during the night. The predictive accuracy of both datasets combined is lower than that of C_1, because the absence of reports for day from C_2 will make most estimators believe that the demand gap between day and night is smaller than the real one.

It is thus clear that, depending on the particular characteristics of different datasets, mixing data may or may not be beneficial.

2.3.2 Computing the Shapley value. Unfortunately, the Shapley value has also been proven to be NP-hard for many domains [10]. The research community has long been using different methods to approximate Shapley values [11, 20, 35]. Some authors have used Monte Carlo to approximate the Shapley value and provide insights into what data is more valuable in ML tasks [18].

In this paper we have tested truncated and non-truncated versions of Monte Carlo (MC), Random Sampling (RS), and Structured Sampling (SS) algorithms. Having evaluated the above approximation algorithms extensively (see details in appendix A) in terms of precision and robustness vs. computational time, we have selected an *ad hoc* tweaked version of the Truncated Structured Sampling (TSS) algorithm since it clearly outperformed the rest of them, and achieved the best trade-off on all the datasets we used for testing: it is able to approximate payoff distributions based on Shapley values with an error of less than 10%, which we consider sufficient for this purpose, in $O(|N|)$ to $O(|N|^2)$ computation time.

2.4 Simpler heuristics for value estimation

One might initially think that the value of the data coming from a provider n_i is given by its volume. In fact, data marketplaces often establish the price of their datasets proportionally to their volume. Hence, we will also consider value distribution based on data volume, which results in the value assignment metric $\mu(n_i) = |S_{n_i}|$, where $|S_{n_i}|$ stands for the data volume of source n_i, or the number of data points originating from this particular source.

The *Leave One Out* (LOO) method, widely used in ML, considers that the value of a source n_i is the difference in performance when the data corresponding to that particular source is removed from the training set. Hence, we define the LOO value of source n_i as $LOO(n_i) = v(S_N) - v(S_{N-\{n_i\}})$, which can be computed in $O(|N|)$ time. In accordance with Eq. 1, the value assignment method in this case, is provided by $\mu(n_i) = LOO(n_i)$.

We also apply specific heuristics tailored to each use case. For predicting demand, we compute the similarity of an input S_K to the aggregate input S_N using the similarity metric v. This tests how close the shape of a each dataset is to the aggregate demand, and is often used in direct detection of time series outliers [19]

For predicting travel time, we apply Shannon's concept of entropy to the histogram of values a dataset provides for relevant spatio temporal features of our model. We define the entropy of a feature j of dataset S_k as $H_j(S_K) = -\sum_{x \in \mathcal{X}_j} p(x) \cdot log(p(x))$. Entropy measures the amount of information or surprise in the values of feature j of the dataset, and requires $\mathcal{X}_j$ to be a discrete set of values. In particular, we calculate three temporal entropy values by discretizing timestamps to their month, weekday and hour, and a spatial entropy by clustering observations and grouping those that are close in space.

We will address the question of how appropriate these heuristics are for estimating the value of data in the next sections.

3 COMPUTING THE VALUE OF COMPANY DATA FOR PREDICTING DEMAND

We start with the case that different companies pool together their data to improve their demand forecasts, either for their own use, or to sell it to an external buyer. In both cases, it is relevant to know how important the contribution of each taxi company is.

3.1 Description of the setting and assumptions

For this purpose, we will focus on metropolitan vehicle-for-hire markets and we will assume that i) service demand observations will be taxi rides reported in a certain spatial coordinates at a certain time, and ii) data sources will be the databases of taxi companies that contain a log of such taxi rides. Our objective will be to forecast the aggregated demand in a control period (T_c) taking as an input the demand reported in an observation period (T_o). For that purpose, we use a multi-seasonal SARIMA algorithm with hourly, daily and weekly sub-components, trained with data over T_o, and producing a forecast ($\hat{S}_K$) over T_c which is compared to the ground truth S_N. Increasing the accuracy of such a prediction model is important both for operational needs (e.g., knowing where to dispatch drivers in anticipation of demand) and planning issues (e.g., deciding where to place taxi service points).

In order to compute results for a real scenario, we will make use of a public dataset of taxi rides from the city of Chicago,[1] which is a log of taxi rides that licensed companies report to local regulatory bodies. We will filter data for the first nine months of 2019 for the analysis (some 11 million rides by 6,469 cars). We will consider the demand for the main 15 taxi companies in that city, plus an additional hypothetical 16$^{\text{th}}$ company, where we aggregate the information from the rest of companies, which account for less than 5% of the total demand.

We will start our analysis by first checking the cases that make collaboration between companies meaningful. For those cases, we will then compute a fair measure of the importance of each individual company based on the quality of the data it offers. We will look at those two matters at both city level, as well as independently for each of the 77 different administrative areas (hereinafter, districts) in which Chicago is divided.

3.2 Demand forecasting at city level

The demand prediction that each company is able to produce on its own yields, in general, an accuracy above 96% at city level as shown in Tab. 1. This means that all companies have enough data to independently predict the future demand with at most a 4% maximum error. Granted that all companies have sufficient data to perform demand prediction accurately on their own, the incentives for collaboration via pooling their data together are very small.

Table 1: City-wide accuracy by company.

Co	Accuracy	Co	Accuracy	Co	Accuracy
All	0.9833	C5	0.9736	C11	0.9659
C0	0.9686	C6	0.9800	C12	0.9845
C1	0.9835	C7	0.9804	C13	0.9725
C2	0.9794	C8	0.9797	C14	0.9767
C3	0.9737	C9	0.9861	C15	0.9724
C4	0.9801	C10	0.9829		

3.3 Demand forecasting at district level

We carried out a similar analysis by isolating the rides of each of the 77 districts of Chicago. Estimating the future demand in this case becomes more challenging and, as we will show soon, often requires collaboration between different companies.

Figure 3 shows the relationship between the forecast accuracy and the number of rides reported within a district. Not surprisingly,

[1] see https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew, last accessed Feb 2022

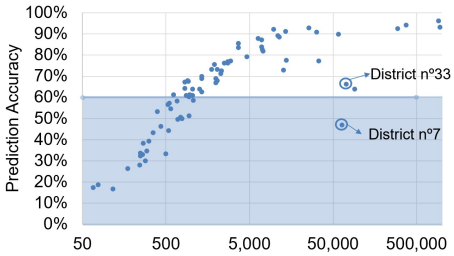


Figure 3: Accuracy improvement by cooperation at district level.

we see that the accuracy is higher in districts with a higher number of reported rides. Predictions at district level are more susceptible to irregular local events than city-wide predictions. For instance, despite being one of the districts with the highest number of reported rides, district number 7 (Lincoln Park), appears to be an outlier in terms of accuracy in Fig. 3. While analysing manually the dataset we found out that a large number of the reported rides were due to a one time event – a James Bay concert at the Riviera Theater, on March 19th. The resulting irregular spike that evening largely explains why the forecasting accuracy remains lower than other districts with smaller volume of demand but more regular patterns.

Another interesting case is district 33 (Near South Side), where the NFL Stadium, McCormick Place and different Museums and city attractions are located. Even though it is reporting a reasonably high number of rides (70k, ranked the fifth district in the city in terms of number of rides), the forecasting algorithm is unable to produce a prediction of high accuracy (goes up to 66% accuracy even with all the available information used). This is due to the event-driven nature of demand in this area, which is not captured by the assumed SARIMA algorithm.[2]

Out of the 77 districts, the forecasting algorithm is able to achieve an accuracy above 60% for 50 of them (those above the shaded region in Fig. 3). This means that even by aggregating all the information available, the particular forecasting algorithm would not be able to predict the future demand with sufficient accuracy for 27 districts.

In order to check whether our findings at city level still hold at district level, we execute the forecasting algorithm in each of these 50 districts for all 16 companies. Figure 4 depicts box-plots (over companies) of the forecasting accuracy improvement from collaboration (Y-axis) in each district (X-axis). Districts are sorted in descending order with respect to the total number of reported rides. We also include city-wide results at the leftmost point of the plot. We find that there is always at least one company that is able to build a forecast model on its own which is very close to the one built by using all the data available. It is not necessarily always the same company across all districts, neither always the biggest one. Smaller companies tend to benefit more from cooperation, which is also less likely to be beneficial in the most popular districts, meaning those that report a large number of rides. However, as we move to smaller districts, the benefits of collaboration start increasing. It is at such areas where it makes sense for different taxi companies to pool their data together to achieve a more accurate forecast.

Focusing on the districts where collaboration makes most sense, we will now compute the relative importance of the data that each company brings, via the notion of the Shapley value.

[2]Areas like this may be amenable to a better prediction accuracy by more complex forecasting algorithms using contextual information but this goes outside the scope of this paper since our focusing is on judging the importance of different datasets for a (reasonable) predictor as opposed to designing the best predictor possible.

Table 2: Shapley value, LOO and nº rides (Rd%) for three districts.

Co	15			17			19		
	SV	LOO	Rd(%)	SV	LOO	Rd(%)	SV	LOO	Rd(%)
1	11.2	0.5	2.5	14.0	0.6	8.3	2.0	0.0	3.4
2	1.8	-0.1	0.8	0.0	-0.1	0.5	1.5	0.0	0.5
3	1.0	0.0	0.3	0.2	0.0	0.5	0.3	0.0	0.0
4	0.4	-0.1	0.2	0.2	0.0	0.0	0.4	0.0	0.1
5	2.3	-0.1	0.9	0.4	0.0	0.5	0.7	0.0	0.8
6	16.4	-1.2	37.9	28.0	8.7	56.2	24.1	3.3	38.6
7	1.1	-0.3	0.4	0.2	0.0	0.4	0.2	0.2	0.5
8	1.1	-0.1	0.8	0.3	0.4	1.4	1.5	0.2	0.5
9	-0.3	0.0	0.2	0.0	0.0	0.2	-0.6	-0.1	0.3
10	2.3	0.4	1.4	0.2	-0.2	0.8	0.9	0.1	0.7
11	0.6	0.0	0.3	0.2	0.1	0.5	1.4	0.0	0.5
12	4.4	-0.1	1.9	0.4	0.1	0.9	2.4	0.1	1.9
13	17.9	0.8	18.1	0.3	-0.2	1.3	4.3	0.0	1.3
14	16.7	-0.9	34.0	17.2	0.0	27.6	26.4	1.9	50.4
15	0.4	0.0	0.1	0.8	0.0	0.3	0.4	0.0	0.1
16	0.2	-0.1	0.2	0.0	0.0	0.8	2.4	0.1	0.5

3.4 Computing the value of information at district level

For the 26 districts marked with an asterisk in Fig. 4, taxi companies would benefit from an increase in forecasting accuracy by combining their data. For each one of them we have computed the Shapley value of the 16 companies using the Shapley formula from Eq. 2. To do that we used cosine similarity as the value function, a test dataset obtained by combining the taxi ride data of all companies active in each district, and the output of the SARIMA model, once trained on the taxi ride data from a particular coalition, as the prediction.

Table 2 summarizes the Shapley value, the LOO value and the percentage of rides reported by each company in the first 3 districts. Figure 5 shows the relationship between the number of rides and the Shapley value for our forecast at district level, which represents the average marginal contribution of its data to the obtained forecast accuracy for that district. Each point in the plot represents a company in one of the 26 districts.

Observing Tab. 2 one may see that different taxi companies can have Shapley values that differ by several orders of magnitude within the same district. Also, the Shapley value of a given company may vary from district to district by a factor of more than ×10 in some cases (see, for instance, companies 1 and 13). Some companies have negative Shapley values in certain districts, meaning they bring *on average* a negative contribution (i.e., they reduce the forecast accuracy) to the coalitions they join.

From Fig. 5 we see that the Shapley values of companies do not correlate well with their number of rides. In fact, the Shapley value for small companies tends to be higher than their corresponding percentage of rides, whereas it is the opposite for large companies. In other words, if we approximated the importance of companies just by the volume of data (rides) they contribute, we would be rewarding large companies, at the expense of smaller ones.

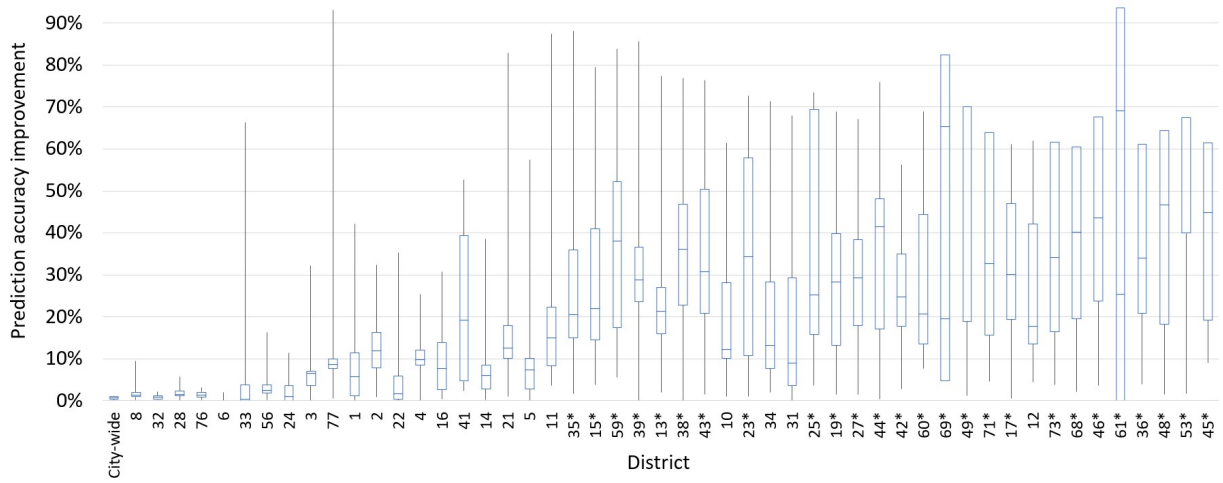


Figure 4: Potential prediction accuracy improvement by cooperation at district level.

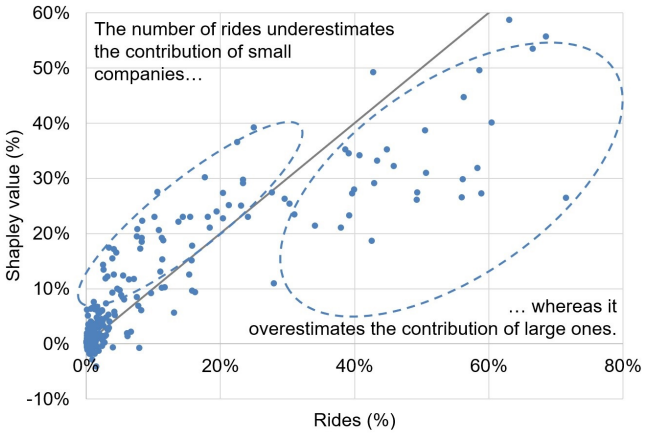


Figure 5: Shapley value vs. nº rides reported by companies for a sample of districts. Each point represents a company in a district.

Similarly, LOO values are weakly correlated ($R^2 = 0.38$) with Shapley values at the level of districts. Notice that even some companies reporting many rides have negative LOO (e.g., company 6 in district 15). As with the case of number of rides, were someone to split an accepted bid based on LOO values, the allocation of payments would deviate significantly from what a splitting based on Shapley would produce.

3.5 Summary

Predicting demand at city level does not require collaboration between different taxi companies since each one can independently estimate city-wide demand. However, when attempting to estimate demand at district level, different companies need to combine their data if they are to achieve a high prediction accuracy. In these cases, neither the data volume a company is providing nor its LOO value reflect accurately its contribution to achieving a better forecast of future demand as given by its corresponding Shapley value.

4 COMPUTING THE VALUE OF INDIVIDUALS DATA FOR PREDICTING DEMAND

In the previous section we applied methods for estimating the value of aggregate data held by taxi companies. In this section we will go a step further, and apply approximations to Shapley value for estimating the value of data held by individual drivers. Recently, some "retail" data marketplaces have appeared that allow individuals or individual IoT sensors to sell their data [8]. These will face additional challenges in terms of scalability to compute the Shapley value over hundreds or thousands of taxi drivers.

4.1 City-wide results

We have computed a TSS Shapley value approximation for a set of $|N| = 4968$ taxi drivers that provided service in Chicago during March and April 2019. In this way we computed the contribution of each driver's data to the forecasting accuracy achieved by the model in predicting the demand in the second half of April using taxi rides from the previous six weeks for training.

In the same way that we proceeded in the wholesale use case, we compared the Shapley value with the number of rides reported by each driver. Figure 6 shows a plot of these two metrics across all drivers. We see that there is no clear relationship between them ($R^2 = 0.1774$), and there seems to be other factors affecting Shapley value. For example, the value of a tuple S of drivers ($v(S)$) seems to be more correlated to the similarity of the input signal to the average ($R^2 = 0.6736$) than to the number of rides that drivers in S are reporting, as suggested by our toy model in Sect. 2. Another interesting finding is that it takes a very small number of drivers to estimate the city-wide aggregate demand. With 7 randomly selected drivers, on average, we can reconstruct the shape of the demand at city level with a 95% accuracy.

4.2 Results at district level

As we just saw, it is possible to build accurate demand forecasts at city level using only a very small number of drivers. We will check whether this also holds for demand forecasts at the district level.

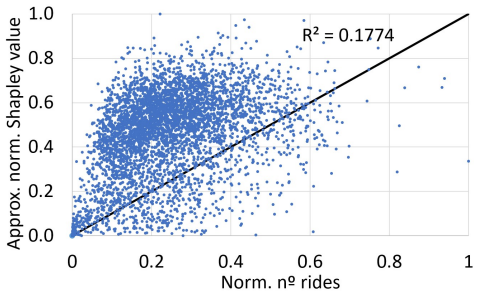
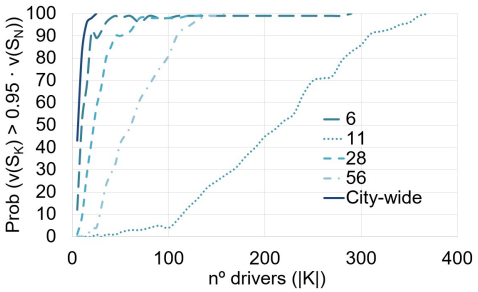
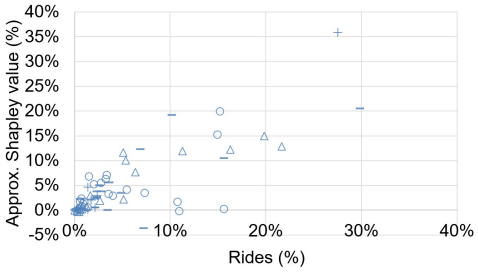


Figure 6: Approx. Shapley value vs. nº of rides across drivers.

For that purpose, we will first quantify the number of necessary drivers, and then proceed to compute the relative value of each driver's data. Figure 7a shows the probability that using a number of drivers indicated in the X-axis one can achieve a prediction accuracy at least 95% of that achieved when using information from all the drivers. Different lines correspond to districts with high (28), medium (6 and 56) and small (11) demand for taxi rides.



(a) Prob ($v(S_K) > 95\% \cdot v(S_N)$ vs nº drivers in K.



(b) Approx. SV vs. nº of rides for tuples K of $|K| = 20$ drivers with $v(S_K)$ exceeding $95\% \cdot v(S_N)$.

Figure 7: Results at district level for individual taxi drivers.

The plot shows that whereas for forecasting city-wide demand, or demand in large districts, few drivers suffice, forecasting the demand of medium-sized and smaller districts requires information from many more drivers. This can be understood by noting that in large districts, the aggregate demand is much more predictable since it is the result of the aggregation of large numbers of independent variables (people that may need a taxi ride). Such demands are known to be easier to forecast, but achieving high forecasting accuracy in medium and small districts requires using the data from tens if not hundreds of drivers. Computing the actual Shapley value is impractical for such numbers of players, but it can be approximated by using the truncated structured sampling approximation discussed earlier in Sect. 2.3.2.

We computed the Shapley values for smaller sets of drivers whose data achieve an accuracy very close to $v(S_N)$ when combined. Figure 7b shows a scatter plot of the approximate Shapley value (Y-axis) vs. percentage of reported rides (X-axis) for a number of such sets of drivers in district 28. Each point represents a driver, and drivers from the same set are represented with the same marker. As observed earlier at city level, the real value of a driver may be very different from that predicted by its number of rides.

5 COMPUTING THE VALUE OF DATA IN PREDICTING DEMAND IN NYC

We have repeated the analysis using a dataset of taxi rides in New York City from April to May 2019.[3] The dataset includes more than 65 million rides from 33 companies in 261 districts.

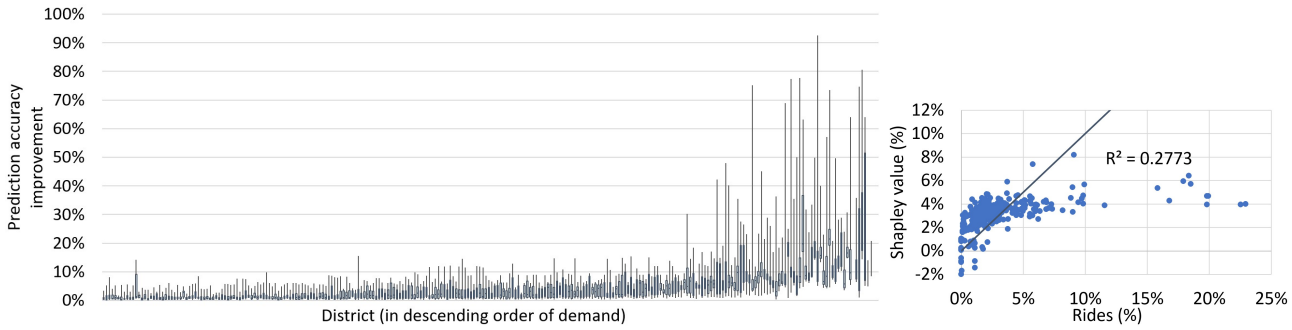
The conclusions from NYC are similar to the ones we drew in detail for Chicago. Figure 8a shows a box-plot (over companies) of the prediction accuracy improvement (Y-axis) by district (X-axis) in the case of NYC. More than 75% of taxi companies are able to predict demand with an accuracy of above 0.9 in 219 districts. Cooperation can improve by at least 10% the accuracy of individual forecasting for more than 50% of the companies in 20 of the smallest districts. There are 4 districts with very few rides in which the forecasting algorithm cannot achieve a high accuracy even combining the data from all companies. In the districts where cooperation between companies made sense, the number of rides reported by each company is again weakly correlated with its Shapley values (R^2 ranging from 17% to 40%, 27% on average, see Fig. 8b), and the same holds for the LOO value. In conclusion, repeating the analysis for a second large dataset verified all our main conclusions obtained from the analysis based on the Chicago dataset.

6 COMPUTING THE VALUE OF DATA IN PREDICTING TRAVEL TIME IN PORTO

As a third and completely different use case, we measured the relative value of the information provided by ubiquitous sensors placed in cars for predicting the travel time between two points of a city. For that purpose, we used a dataset[4] that provides the routes of 448 taxis within the city of Porto from Jul'13 to Jun'14, and a Random Forest regressor model aimed at predicting how long it takes to get to a destination from any other point in the city. We trained this model with the UTM coordinates of the start and end points, the time of the year (month, weekday, hour), the type of day (i.e., workday, holidays or day before holidays) and the travel distance. Finally, we ran two experiments to predict travel time to the airport and to São Bento train station as key points of interest.

A first analysis reveals substantial differences in the amount of data provided by each individual car. Figure 9 shows a histogram of the number of rides reported by each taxi (in grey bars), and a scatter plot comparing the percentage they represent in the whole dataset (X-axis), and in the subset of rides ending in the corresponding

[3] see https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page, last access Feb'22
[4] see https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/data, last access Feb'22



(a) Box-plot (over companies) of potential prediction accuracy.

(b) Shapley value vs. nº rides in small districts.

Figure 8: Area-level results by company for demand prediction in NYC.

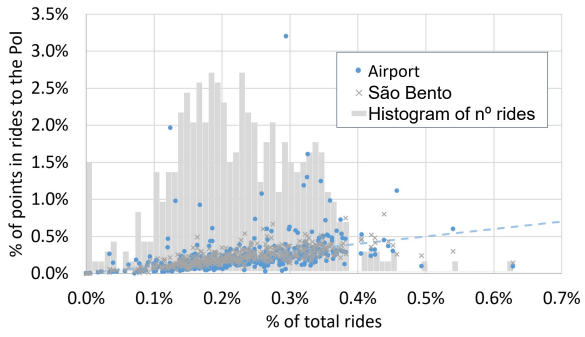


Figure 9: Total vs nº rides to Porto's airport and to São Bento St.

point of interest (Y-Axis). Not only do vehicles provide different volume of information, but they also exhibit different behavior. For example, the one reporting the largest number of rides, hence more to the right, turns out to be reporting only a few rides to the airport and the train station. On the contrary, some drivers that appear well above the dotted line travel much more frequently to the airport than the average.

We randomly split input data between train (80%) and test (20%) set (S_{test}). Using all the information available in the training set, the models provide reasonable results (an average R^2 score above 0.8 for both destinations). We calculate the value that data of each individual vehicle is bringing to the prediction model by comparing its predictions to the real travel time in the test set using the R^2 score as the output evaluation metric v.

Even though PIMS usually reward data of individuals equitatively, our experiments show that each taxi brings very different value to the model. Whereas the Shapley value of some vehicles is close to twice the average, others account for a negligible or even negative contribution. Some of them are valuable for predicting the travel time to the airport, but not so valuable for predicting the travel time to the São Bento train station.

It is also quite common to set the price of data based on its volume [8, 27], assuming that the more points reported, the more valuable a piece of data would be to the model. However, as the correlation plot in Fig. 10a shows, this is not exactly the case. The Pearson correlation (R) of Shapley values for predicting travel time with the number of rides reported is 0.56 and 0.39 respectively. Such results evidence that the more rides a driver reports the more likely it is that its data will improve the accuracy of the model. However, two individuals reporting the same number of rides may bring a very different value to the model, which supports the hypotheses that other complex dependencies and factors must be considered in this calculation.

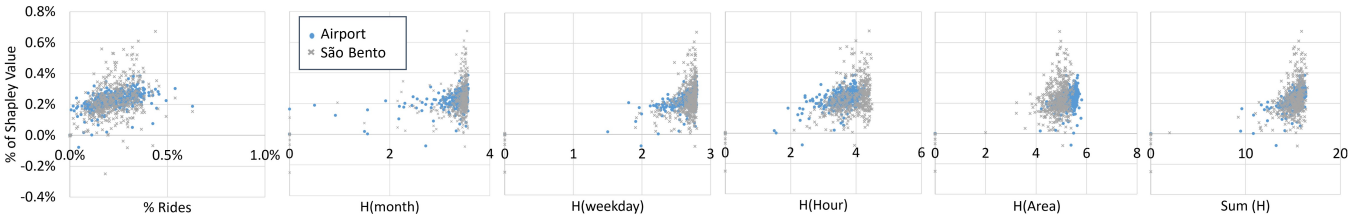
We also considered the the spatio-temporal *diversity* of data and its relationship to data value. We want to check whether wider time span and the more dispersion across space lead to higher value for forecasting. We have measured this *diversity* through the concept of information entropy. It turns out that the entropy rates of spatial and temporal features show significant correlation with Shapley values (R ranging from 0.45 to 0.69), as Figs. 10(b-f) show. Therefore the higher the entropy of information of an individual taxi, the more *diverse* in terms of time and space, and hence the more valuable for the prediction algorithm it will be. On the contrary, taxis failing to provide data for specific days of the week, those that only report rides at certain hours or in specific areas, or those that stopped their activity for several months apparently mislead the prediction algorithm and, hence, are less valuable. This result points to another potential approach for approximating the value of data without having to compute the Shapley value, but instead relying on the above mentioned diversity feature.

Finally, as Fig. 11 shows, LOO values are small ($|LOO| < 0.01$) and uncorrelated with the Shapley value or with the number of rides. As a result, any reward attribution or ranking of data sources based on LOO would be close to random.

7 RELATED WORK

The use of spatio-temporal data in transportation and smart city applications has attracted much attention from the research community. Different works look at how knowledge extraction from spatio-temporal data can significantly improve the effectiveness of transportation [38], mobility prediction [5], or last mile delivery [14], among others.

Some authors have also studied the *intrinsic* value of spatio-temporal information, and calculate it as the reduction of the uncertainty about the position of an individual [25]. Similar to ours, other works have dealt with the *extrinsic* value of spatio-temporal data for a specific problem in a specific context [3, 4], while others



(a) $R_1 = 0.56, R_2 = 0.39$ **(b)** $R_1 = 0.63, R_2 = 0.48$ **(c)** $R_1 = 0.67, R_2 = 0.47$ **(d)** $R_1 = 0.69, R_2 = 0.57$ **(e)** $R_1 = 0.62, R_2 = 0.45$ **(f)** $R_1 = 0.69, R_2 = 0.60$

Figure 10: Pearson correlation of Shapley values with data features for predicting travel time to Porto's airport (R_1) and São Bento Station (R_2).

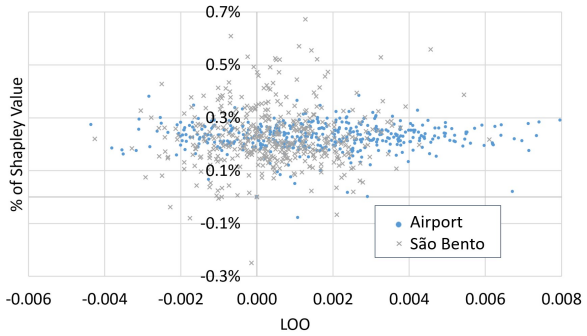


Figure 11: LOO vs Shapley values for predicting travel time in Porto.

have introduced the notion of privacy in pricing spatio-temporal data [24]. These valuable works differ from ours in 1) they adapt different notions of value instead of using the more generic Shapley value used in our work, 2) they look at different applications domains (e.g, location-based marketing, heavy traffic prediction) than the ones we study, and 3) they deal with different problems, such as identifying the best moment for bidding or acquiring a new data point, but they do not address the one of rewarding data sources contributing to a single dataset based on their value.

The idea of providing *micropayments* to users for their personal data has received a lot of public attention [21, 28]. Finding scalable and fair ways to compute value-based contributions to a ML problem is key in calculating the contribution of individuals to the data economy will require to find fair scalable ways to do it. More recent work describes fundamental technological challenges that need to be addressed for the above vision to be fulfilled [22].

Previous works have already used the Shapley value to measure the utility of training data for different other ML problems [18, 30]. Another body of related work has to do with computational aspects of the Shapley value across different application domains. Several works have looked at computational aspects of Shapley value and have developed efficient exact and approximation algorithms for particular types of problems such as image classification, graph centrality, and others [11, 18, 20, 35, 40]. Recent works propose using appraisal functions adapted to the current model, and introduce multi-party computation to evaluate and select private training data from a marketplace [37]. None of the above works has looked at valuation issues relating to spatio-temporal data.

Finally, some survey papers have studied the way commercial data marketplaces [31, 33, 34] work, including our recent work which is, to the best of our knowledge, the largest survey on this subject matter, and the first one studying the business models of PIMS [8]. Moreover, we recently carried out a first of its kind measurement study on the prices of data in commercial data marketplaces [6]. The research community is also actively working on proposing novel architectures for data marketplaces to train ML models provided by buyers with data from sellers [2, 13, 16, 26]. Despite their efforts, we are not aware of any commercial marketplace that has implemented such value-based data trading proposals, yet.

8 CONCLUSION AND FUTURE WORK

In this work we have looked at the problem of how to compute the relative importance of different spatio-temporal datasets that are combined in order to improve the accuracy of demand and travel time forecasting for taxi rides in large metropolitan areas such as Chicago, Porto and New York. Our main result has been that the importance of each dataset differs and cannot be deduced via simple heuristics based on data volume or leave-one-out methods, but instead one needs to look deeper and consider the complex ways in which different datasets complement one another, which is what the Shapley value does. This applies when combining data from entire companies, but as well for data from individual drivers.

However, the Shapley value has inherent scalability issues and, even using efficient approximation algorithms, requires a $O(N^2)$ computation time which is still insufficient for thousands of taxis or complex prediction models. We have also noticed the correlation between data volume and its Shannon information entropy with the value of spatio-temporal data. Were Shapley values based on functions combining volume and entropy able to approximate those based on repetitively training a complex ML model, they would arguably be much more efficient in terms of computing time.

We have addressed our main question of whether simple or more complex methods need to be used in order to split the value of a dataset in a fair manner in a "full information" setting, i.e., assuming that we can test the forecasting potential of each individual dataset or coalition of datasets. Designing a fully functioning marketplace that implements our ideas is a much bigger problem that we are currently working on. Several additional challenges remain to achieve this. For example:

- Data buyers need to have a way to estimate the value of a coalition of datasets - either for their ML algorithm or for a given data valuation function - without, however, having access to raw data that they have not purchased yet. This

can be done in different ways, including having the marketplace execute the training locally and communicating to data buyers the achieved accuracy, or using secure multi-party computation [37].
- Data buyers also need to be protected against strategic data sellers that may modify their datasets artificially, or just sell replicas of the same data, in order to receive higher rewards.
- Once a bid has been accepted, splitting it in proportion to the different Shapley values is a reasonable and fair approach but need not be the only one. To attract new sellers, and to retain existing ones, a marketplace may choose to pay a minimum amount to some sources, regardless their Shapley values. Other considerations may also prompt it to partially deviate from splitting according to Shapley value.
- Partial information models, in which not even the marketplace has full information on each and every dataset, can also be considered.

Last, we are working at applying or developing practical methods to compute the value of data for other tasks related to spatiotemporal data, such as predicting congestion or forecasting optimal routes, and for tasks related to other types of data. This includes optimizing Shapley approximation algorithms, measuring Shapley for simplified models or, leveraging the results we obtained in this paper, designing *ad hoc* valuation heuristics based on the most important features to identify suitable data for improving the performance of a model.

9 ACKNOWLEDGEMENTS

This paper has been funded by the European Union's Horizon 2020 Research and Innovation program under the DataBri-X (Project No. 101070069) and the PIMCITY (Project No. 871370) projects.

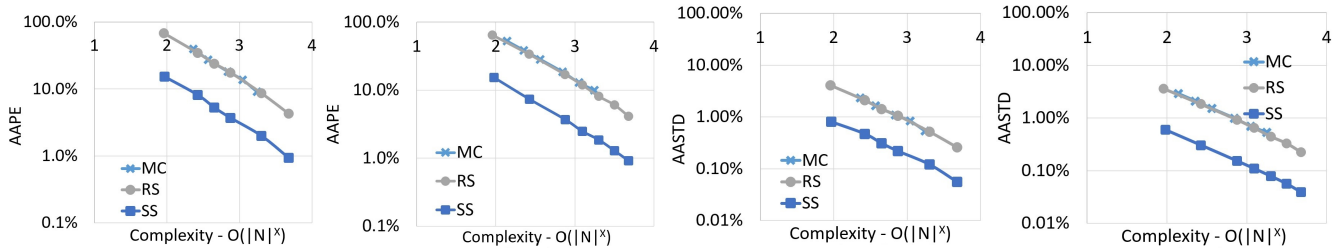
REFERENCES

[1] Advaneo. 2022. https://www.advaneo-datamarketplace.de/. Last accessed: Jun'22.
[2] A. Agarwal, M. Dahleh, and T. Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In *Proc. of ACM EC'19*.
[3] H. Aly, J. Krumm, G. Ranade, and E. Horvitz. 2018. On the Value of Spatiotemporal Information: Principles and Scenarios. In *Proc. of ACM SIGSPATIAL '18*. ACM.
[4] H. Aly, J. Krumm, G. Ranade, and E. Horvitz. 2019. To Buy or Not to Buy: Computing Value of Spatiotemporal Information. *Proc. of ACM SIGSPATIAL '19*.
[5] L. Amichi, A. C. Viana, M. Crovella, and A.F. Loureiro. 2021. From Movement Purpose to Perceptive Spatial Mobility Prediction. *In Proc. of SIGSPATIAL '21*.
[6] S. Andrés, C. Iordanou, and N. Laoutaris. 2021. What Is the Price of Data? A Measurement Study of Commercial Data Marketplaces. (2021).
[7] S. Andrés and N. Laoutaris. 2020. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. arXiv:2012.08874
[8] S. Andrés and N. Laoutaris. 2022. A Survey of Data Marketplaces and Their Business Models. *SIGMOD Record* (2022).
[9] International Data Spaces Association. 2022. IDSA. The future of the data economy is here. https://internationaldataspaces.org/. Last accessed: Jun'22.
[10] Y. Bachrach, E. Elkind, R. Meir, D. Pasechnik, M. Zuckerman, J. Rothe, and J. Rosenschein. 2009. The Cost of Stability in Coalitional Games. In *Algorithmic Game Theory*. Springer Berlin Heidelberg.
[11] S. Cabello and T. M. Chan. 2022. Computing Shapley values in the plane. *Discrete & Computational Geometry* (2022).
[12] Javier Castro, Daniel Gomez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers and Operations Research* 36 (05 2009), 1726–1730. https://doi.org/10.1016/j.cor.2008.04.004
[13] L. Chen, P. Koutris, and A. Kumar. 2019. Towards Model-Based Pricing for Machine Learning in a Data Marketplace. In *Proc. of SIGMOD'19*. ACM.
[14] C. F. Costa and M. A. Nascimento. 2021. Last Mile Delivery Considering Time-Dependent Locations. In *Proc. of SIGSPATIAL'21*. ACM.
[15] S. S. Fatima, M. Wooldridge, and N. R. Jennings. 2008. A linear approximation method for the Shapley value. *Artificial Intelligence*.
[16] R. C. Fernandez, P. Subramaniam, and M. J. Franklin. 2020. Data Market Platforms: Trading Data Assets to Solve Data Problems. *Proc. VLDB Endow.* 13, 12 (2020).
[17] Gaia-X. 2022. https://gaia-x.eu/. Last accessed: Jun'22.
[18] A. Ghorbani and J. Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning.
[19] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han. 2014. Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 26, 9 (2014).
[20] R. Jia, D. Dao, B. Wang, F. Hubis, N. Gurel, C. Zhang, C. Spanos, and D. Song. 2019. Efficient task-specific data valuation for nearest neighbor algorithms. *Proc. of VLDB Endow.* (2019).
[21] J. Lanier. 2013. *Who Owns the Future?* SIMON and SCHUSTER.
[22] N. Laoutaris. 2019. Why Online Services Should Pay You for Your Data? The Arguments for a Human-Centric Data Economy. *IEEE Internet Computing* (2019).
[23] S. Mehta, M. Dawande, G. Janakiraman, and V. Mookerjee. 2019. How to Sell a Dataset?: Pricing Policies for Data Monetization. *Information Systems Research*.
[24] K. Nguyen, J. Krumm, and C. Shahabi. 2020. Spatial Privacy Pricing: The Interplay between Privacy, Utility and Price in Geo-Marketplaces. In *Proc. SIGSPATIAL '20*.
[25] K. Nguyen, J. Krumm, and C. Shahabi. 2021. Quantifying Intrinsic Value of Information of Trajectories. In *Proc. of SIGSPATIAL'21* (Beijing, China). ACM.
[26] O. Ohrimenko, S. Tople, and S. Tschiatschek. 2019. Collaborative Machine Learning Markets with Data-Replication-Robust Payments. (2019). arXiv:1911.09052
[27] J. Pei. 2020. Data Pricing – From Economics to Data Science. In *Proc. of ACM SIGKDD'20*.
[28] E. Posner and G. Weyl. 2018. *Radical Markets. Uprooting Capitalism and Democracy for a Just Society*. Princeton Univ. Press.
[29] Ocean Protocol. 2022. https://oceanprotocol.com/. Last accessed: Jun'22.
[30] B. Rozemberczki, L. Watson, P. Bayer, H. Yang, O. Kiss, S. Nilsson, and R. Sarkar. 2022. The Shapley Value in Machine Learning. arXiv:2202.05594
[31] F. Schomm, F. Stahl, and G. Vossen. 2013. Marketplaces for Data: An Initial Survey. *SIGMOD Record* 42, 1 (May 2013), 12 pages.
[32] L. S. Shapley. 1952. A Value for n-Person Games. (1952).
[33] M. Spiekermann. 2019. Data Marketplaces: Trends and Monetisation of Data Goods. *Intereconomics* (2019).
[34] F. Stahl, F. Schomm, L. Vomfell, and G. Vossen. 2017. Marketplaces for Digital Data: Quo Vadis? *Computer and Information Science* 10 (2017).
[35] S. Touati, M. S. Radjef, and L. Sais. 2021. A Bayesian Monte Carlo method for computing the Shapley value: Application to weighted voting and bin packing games. *Computers and Operations Research* (2021).
[36] T. van Campen, H. Hamers, B. Husslage, and R. Lindelauf. 2017. A new approximation method for the Shapley value applied to the WTC 9/11 terrorist attack. *Social Network Analysis and Mining* 8, 1 (02 Dec 2017).
[37] X. Xu, A. Hannun, and L. Van Der Maaten. 2022. Data Appraisal Without Data Sharing *(Proc. of Machine Learning Research)*.
[38] L. Yan, H. Shen, Z. Li, A. Sarker, J. A. Stankovic, C. Qiu, J. Zhao, and C. Xu. 2018. Employing Opportunistic Charging for Electric Taxicabs to Reduce Idle Time. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2018).
[39] T. Yan and A. D. Procaccia. 2021. If You Like Shapley Then You'll Love the Core. *Proc. of the AAAI Conf, on Artificial Intelligence* 35, 6 (May 2021), 5751–5759.
[40] K. Zhao, S. H. Mahboobi, and S. Bagheri. 2018. Shapley Value Methods for Attribution Modeling in Online Advertising. (04 2018).

A TESTING DIFFERENT SHAPLEY VALUE APPROXIMATION ALGORITHMS

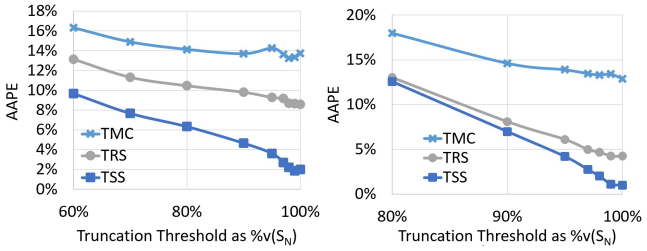
We tested Monte Carlo (MC) [18], Random Sampling (RS) [12], and Structured Sampling (SS) algorithms to approximate Shapley values and compared them to the exact Shapley values calculated using Eq. 2. When structuring samples of permutations in SS algorithms, we tailored a solution to address problems where the position of *players* in a permutation strongly determines their marginal contribution [15, 36]. Whereas RS and MC randomly select the permutations to evaluate, SS ensures that all companies or individuals appear the same number of times in each position of the permutations. This we achieve by means of defining and processing the average marginal value of each dataset for samples of r blocks of latin squares of $|N|$ permutations.

First, we carried out a test for approximating the value of companies in two different demand prediction scenarios (city-wide and medium-size district 35) of Chicago (*Scenario 1* or *Sc1*). Then, we repeated the test for approximating the value of individuals in 10 different tuples of 6 to 16 drivers yielding a good accuracy (close to



(a) AAPE vs. complexity Sc1. (b) AAPE vs. complexity Sc2. (c) AASTD vs. complexity Sc1. (d) AASTD vs. complexity Sc2.

Figure 12: Accuracy and robustness vs. complexity of the algorithms



(a) AAPE vs. truncation Sc1. (b) AAPE vs. truncation Sc2.

Figure 13: Accuracy vs. truncation threshold

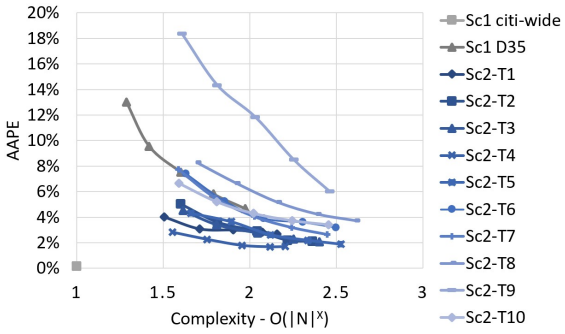


Figure 14: TSS AAPE vs complexity for r = 1, 2, 4, 8, 16.

0.8) in predicting travel time to Porto's airport (*Scenario 2* or *Sc2*). Our cases cover individuals and companies whose Shapley values are similar, and situations in which the value of companies and individual drivers differ significantly (up to more than ×5). Given their stochastic nature, we tested truncated and non-truncated versions of MC, RS and SS algorithms 50 times for each combination of input parameters and tuple of data sources.

We measured the performance of the algorithms in terms of:
- *Accuracy*, through the average average percentage error (AAPE) compared to the real Shapley values.
- *Robustness*, measured as the average average standard deviation (hereinafter, AASTD) of their outputs.
- *Complexity*, measured in terms of the actual number of training-prediction cycles computed in each case. For example, a complexity of 2 means that the algorithm required $O(|N|^2)$ such cycles.

Figure 12 shows a comparison of MC, RS and SS in terms of accuracy and robustness for different levels of complexity (X-axis). We introduce complexity by reducing the maximum variation of Shapley values for the MC model to converge, and by increasing the number of permutations evaluated in RS and SS approximations. In general, the more combinations evaluated, the more accurate and, especially, the more robust the approximations are. As shown in the figure, SS clearly outperforms both RS and MC, and delivers more consistent outputs across executions, which are also much closer to the exact Shapley values. This is thanks to the planning of the sample permutations, which helps in reducing the randomness of the sample, and hence increases the accuracy of the approximation.

Figure 13 shows the effect of truncation on the accuracy for the three algorithms. According to our results, SS is slightly more sensitive to truncation, but it always outperforms the other algorithms. Moreover, it is possible to easily control the trade-off between accuracy and execution time by tuning r and the truncation threshold. We chose to use a *truncation threshold* of $0.95 \cdot v(S_N)$ since it divides the execution time by 4, while maintaining the AAPE below 5%.

To select r. we tested the TSS algorithm for increasingly complex ($r = 1, 2, 4, 8, 16$) approximations. Figure 14 shows the AAPE for the different test tuples of drivers and companies in both scenarios. Overall, we observe that for $r = 2$ and a complexity between $O(|N|)$ to $O(|N|^2)$ the AAPE is 5.2% on average and below 15% in all the cases, which we consider reasonable for distributing payoffs.