Computing the Value of Spatio-Temporal Data in Wholesale and Retail Data Marketplaces

Santiago Andres Azcoitia[a], Marius Paraschiv[b], Nikolaos Laoutaris[b]

[a]*IMDEA Networks Institute, Univ. Carlos III, Leganes (Madrid) Spain*
[b]*IMDEA Networks Institute, Leganes (Madrid) Spain*

Abstract

Spatio-temporal information is increasingly used for driving a plethora of intelligent transportation, smart-city, and crowd-sensing applications. At the same time, different types of data marketplaces are proposed for de-siloing and monetising individual and enterprise data. In this paper we study the problem of estimating the relative value of spatio-temporal data sold in wholesale and retail data marketplaces for the purpose of forecasting future demand in a certain area, *e.g.*a city. Using as case studies large datasets of taxi rides from Chicago and New York, we ask questions such as "When does it make sense for different taxi companies to combine their data?" and "How should different companies be compensated for the data that they share?". We then turn our attention to the even harder problem of establishing the value of the data brought to retail marketplaces by individual drivers. Overall, we show that simplistic approaches, such as assuming that the value of the data held by companies or drivers is proportional to its volume are inaccurate, because they fail to consider the complex complementarities that may exist among different datasets. To remedy this, more complex notions of value-sharing from economics and game-theory, such as the Shapley value need to be used to capture the effect of mixing datasets on the accuracy of forecasting algorithms driven by them. Applying the Shapley value to large datasets from many sources is computationally challenging. We use structured sampling to overcome such scalability challenges and manage to compute accurately the importance of different data sources, even when their number ranges in the thousands, as in the case of all the taxi drivers in a large metropolis.

Keywords: Data value, Shapley value, data marketplace, personal information management systems (PIMS), intelligent transportation

February 2020

1. Introduction

Data-driven decision making is bringing significant improvements to many sectors of the economy, including in 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 infrastructure planning in a city [23, 27, 9, 24, 25], whereas companies like Uber are increasingly deploying and operating sophisticated systems for optimising their operations using live data.[1] In the above and many other settings, data are considered as important corporate assets, next to more traditional ones, such as labour force, capital, and infrastructure. Therefore, it is becoming imperative to be able to measure the value of data, especially when they get combined from multiple sources.

The focus of this paper is on establishing the relative value of different spatio-temporal datasets for forecasting future demand for a service across space and time in a wide area, such as a city, or a metropolitan region. 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, or provisioning service points. It can also be sold as a service to third parties that do not own such data but need it for important business decisions.[2] To fulfill such needs, several data marketplaces have appeared, making available for purchase both *wholesale* enterprise (see IOTA[3] or Airbloc[4]) and *retail* data produced and held by individuals (see Digi.me, MyDex, HAT, EarnieApp, Citizen.me, Meeco or OwnYourInfo [5] for some examples of so called Personal Information Man-

[1] Examples of how Uber leverages spatio-temporal data across its main processes and operations `https://eng.uber.com/forecasting-introduction/` (last accessed February 2020).

[2] Such commercial services are already offered by several telcos, banks, and other enterprises, see for example: `www.business-solutions.telefonica.com/en/products/big-data/business-insights/smart-steps/`

[3] See `https://data.iota.org/`

[4] See `https://airbloc.org/`

[5] See `https://digi.me/`, `https://mydex.org/`, `https://www.hubofallthings.com/`, `https://ernieapp.com/`, `https://www.meeco.me/` or `http://www.ownyourinfo.com/`, last accessed Jan 2020.

agement Systems (PIMS)).

For the purpose of our work, we concentrate on vehicle-for-hire demand prediction in Chicago and New York. While our examples and findings are specific to this particular urban mobility use case, the methods that we develop for assigning value to spatio-temporal datasets held by (taxi) companies and individuals (drivers) 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. Coming back to our own setting, we develop data valuation methods to answer a series of fundamental questions pertaining to both wholesale and retail data fusion. For example, "Does combining multiple datasets of past taxi rides always benefit the forecasting accuracy of future demand?". Also, when it does, "How should we attribute the improved forecasting precision to the individual datasets used to produce it?". Finally, we look at the interplay between fairness and scalability/practicality and ask "Can the value of a spatio-temporal dataset be approximated by its volume alone?".

To answer the above questions, we use the Shapley value [22] from collaborative game theory as a baseline metric for establishing the importance of individual *players* (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 few tenths of companies, but becomes impossible when considering hundreds or thousands of drivers.

A much simpler way to compensate data providers is based on the volume of data, in our case, taxi rides, that they report. While certainly more practical, the latter assumes that any reported ride from the past has equal value for predicting rides of the future. The latter is clearly not the case. For example, in the context of a single data source, a reported ride that helps in completing the picture regarding the multiple periodicity existing in the demand is more useful for a forecasting algorithm than a "one-off" ride that does not relate with any phenomena that are amenable to prediction. Things become even more complex when considering data from multiple sources. In this case, the value of a data point cannot be judged only with respect to other data points from the same datasets but, instead, has to be considered in the broader context given by the complementarities existing among different datasets. For example, a taxi driver or a company that reports high demand

for time periods and locations that are already known to be of high demand has less value for prediction than reports of high demand concerning locations and periods that are not covered by the datasets brought by other sources.

Our contributions: We describe how to apply and adapt the notion of the Shapley value to the problem of establishing the value of different spatio-temporal datasets used in forecasting for-hire transportation demand in a city. The above requires addressing interpretability challenges about what the Shapley value means in that setting, as well as scalability challenges arising when having large numbers of data sources. To the best of our knowledge, ours is the first work using rigorous notions of fairness from economics and game theory to data valuation challenges in the important area of data-driven transportation. We first study wholesale data fusion at the granularity of entire companies. Since the number of such companies covering the same geographical area is typically small, the value of their data can be computed directly from the definition of what the Shapley value is. 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 structured sampling [10] performs much better than other approaches such as Monte Carlo [19, 11] and random sampling.

By applying our model and valuation algorithms to taxi ride data from Chicago and New York, we find that sufficiently large taxi companies hold enough data 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 this case. On the other hand, when the objective is to predict the demand at a finer – district-level – granularity, then there are plenty of districts in which companies have to combine their data in order to achieve a sufficient forecasting accuracy. Computing the relative value of different contributions in such cases, we find that there exist companies whose data value differs by several orders of magnitude. Also the importance of the data of a given company can vary as much as $\times 10$ from one district to another. More interesting, the importance of a company's data does not necessarily correlate with the amount of data that that the company brings to the collaboration, i.e., there are companies that report relatively few rides but have a larger impact on forecasting accuracy than companies that report many more rides. Similar phenomena are observed at the finer level of individual drivers. We show that combining data

from relatively few drivers one can easily detect the hours of peak demand at the city level. At district level, however, more data needs to be combined, and this required making use of our fastest approximations of Shapley value based on structured sampling.

2. Problem Description

In all the data marketplaces mentioned before, deciding a price for a dataset is left to the two trading partners to negotiate. The data seller can set a fixed price, or the data buyer can make a bid that the data seller accepts or rejects [17]. The problem we study in this paper differs substantially from such bilateral negotiations. What we will study is how to assess the relative importance of multiple data sources when they combine their data. The resulting aggregate dataset can be used by the same (federated) data sources or can be sold to third parties. In both cases, federation is creating a surplus that can be monetary (when the data is eventually sold) or be had "in kind" (when the data is used by the federated data sources to improve the quality of their forecasts). In both cases there is a need to compute the relative importance of each source to the achieved collective performance.

Next we introduce some notation that will be used in the rest of the paper. Let S denote a dataset and its utility (or value) to a predictive algorithm be $v(S)$. The notion of *value* is to be interpreted as the ability of an algorithm to produce accurate predictions, when trained on the respective dataset. Thus, our notion of data *value* is linked to an algorithm's estimation *accuracy*. Data may be provided by different *sources*. We denote the set of such data sources by $N = \{n_1, n_2, ..., n_{|N|}\}$.

Consider such a set of data sources, each one reporting a series of past demand observations (x, t), where x is the spatial coordinate of the demand point (for example its latitude and longitude, or the corresponding district in a city) and t is the time at which the demand took place. As such, the data is in the form of a *time series* or *signal* $y(t)$ in an observation period T_o, to which we shall sometimes refer to as a *demand function*. As the name suggests, $y_N(t)$ represents the aggregate demand function for all data points reported by the entire set of providers N.

This aggregate signal is then used to train a forecasting model, a multi seasonal SARIMA in our case, whose output is a time-series $\hat{y}(t)$ in a control period T_c, representing the model's future total demand prediction. For every use case, this value will remain fixed, it is computed once and used as the

ground truth when estimating the value of data coming from subsets of the set of data providers.

To give our value function a mathematical expression, we define the value of the data coming from a subset $K \subseteq N$ of data providers, say S_K, as the cosine similarity between the predicted signal, and the reference aggregate,

$$v(S_K) = \text{CosSimilarity}(y_N(t), \hat{y}_K(t)), \tag{1}$$

where $\hat{y}_K(t)$ is the prediction that the model returns, if trained only on the data set S_K. Defined in this manner, the value function expresses how accurately predictions resulting from a data set can reproduce the total reference aggregate, over a certain *control period*, T_c used for this purpose. The cosine similarity has been chosen among other similarity metrics because we are mostly interested to predict the "shape" of future demand as opposed to its exact magnitude. Having the shapes allows us to identify busy and less busy hours and districts (when the forecasting is repeated at district level). From the shape one can easily extrapolate the actual demand in terms of passengers by scaling up by a factor proportional to the number of taxis used in producing the prediction vs. the total number of taxis in the city. The latter is usually public information in most cities. Figure 1 shows a block diagram that describes the general prediction model used throughout the paper.

Different groups or *coalitions* of providers may achieve different prediction accuracies than others. Given that the value of a coalition of data providers is measured by its forecast accuracy, how should we attribute this accuracy to each one of the data providers? In other words, if we were to pay each individual provider for its training data, how should we compute the resulting payment in a fair manner?

One might initially think to pay data providers based on the volume of data they report. This, however is, as we shall see, not necessarily a good value estimate. Some factors affecting the capacity of data to bring a positive contribution to the estimation of the total demand aggregate are: *spatial and temporal coverage*, as data from a provider which is only active in certain locations cannot offer information regarding the demand in the rest of the city; *redundancy*, there can be a degree of overlap between the data given by one company and another; *complementarity*, a service provider might only be active at night, or during the weekend and consequently might not be able to infer the demand for the entire week by itself. However, combined

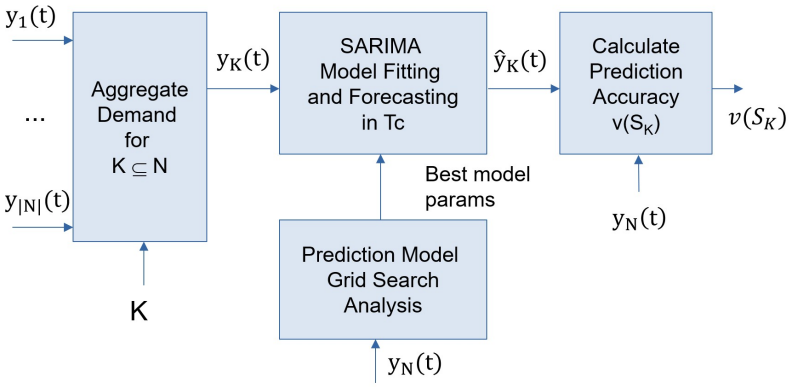

Figure 1: Block diagram describing the general prediction model used throughout the paper. The model constructs the aggregate input (or training) demand $y_K(t)$ from a set of sources $K \subseteq N$ by combining their individual demands $y_i(t)$ during an observation (or training) period T_o. This input then drives a prediction model, which in our case is a multi-seasonal SARIMA algorithm with daily and weekly sub-components. The parameters $(p,d,q)x(p,d,q)^s$ for such a predictor are obtained via grid search analysis to minimize the AIC (Akaike Information Criterion, [2]). As a result, the prediction model produces a forecast $\hat{y}_K(t)$ in the control period T_c which is compared to the real demand as described in Eq. 1.

with the data from another, active during the day or the working part of the week, the aggregate may indeed prove valuable. In the coming sections we demonstrate all of the above using real data, and propose a method for taking into consideration all complex complimentarities that may appear when mixing data from multiple sources.

3. Computing the importance of data in wholesale collaborations

We start with the case that different companies pool together their data in order to improve the forecasting accuracy for their own use, or for selling the resulting aggregate dataset to external data buyers. In both cases it is relevant to know how important the data contribution of each data source is.

3.1. Description of setting and assumptions

For the purpose of this use case, 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

Table 1: Chicago City taxi rides dataset (retrieved during Nov'19). Brief description and statistics for the whole data set and for the specific period which was used in the simulations.

Time period	01-01-2013 - 09-01-2019	01-01-2019 - 09-01-2019
Rides	94 millions	11.1 millions
Companies	160 with 101 individual licenses	58. 94% rides from top 15 companies
Districts	77 districts (administrative communities) of Chicago City	
Taxi Ids	19,014. 55% of the total licenses associated to 5 companies	6,469

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 taking as an input the demand reported in an observation period. 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), so as it would make sense for a company to collaborate in case its prediction accuracy could be significantly improved by pooling similar data with other companies.

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, [6] which is a log of taxi rides that licensed companies report to local regulatory bodies. This dataset consists of more than 94 million rides from 160 companies, spanning from 2013 to 2019. We will filter data for the first half of 2019 for the analysis (see Table 1 for a summary of the properties of this dataset). We will consider the demand for the main 15 taxi companies in that city, plus an additional hypothetical 16th company, where we aggregate the information from the rest of companies, which account for less than 5% of the total demand. In section 5 we will also present the results for a similar data set from New York City.

We will start our analysis by first checking the cases that make collaboration between companies meaningful. For those cases, we will then compute

[6] see `https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew`, last accessed January 2020)

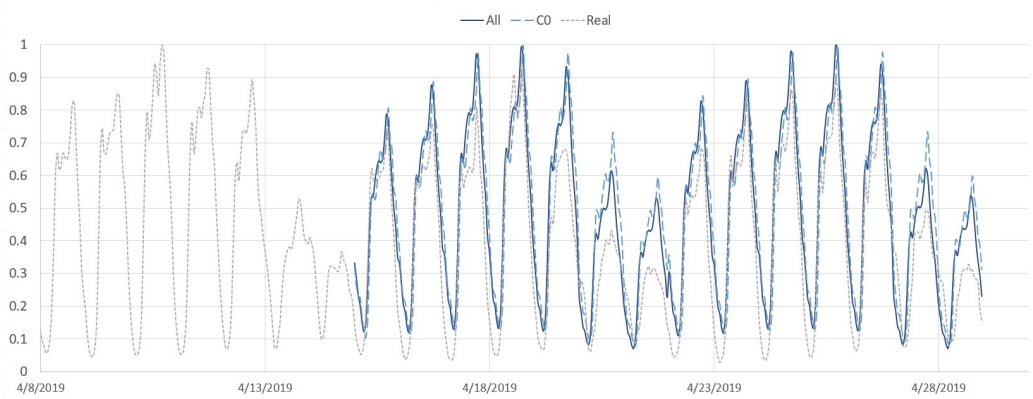

Figure 2: Example plot of a city-wide SARIMA model fit using the information from all companies and only from company labelled as C0

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

Figure 2 shows a prediction sample for a control period between Apr 15th and Apr 28th 2019 based on the observations of the previous weeks. It compares the real observed demand to the predicted demand using information from *all* companies and only from the company labelled as *C0*. Similar plots are obtained for the rest of the companies. Table 2 shows a summary of the forecast accuracy achieved by using all the information available and by using only the information from each company. According to our results, the demand prediction that each company is able to produce on its own yields, in general, an accuracy above 96% at city level. This means that all companies have enough data to independently predict the future demand with at most a 4% maximum average 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.

3.3. Demand forecasting at district level

We performed a similar analysis by isolating the rides of each of the 77 districts of Chicago. Estimating the future demand in this case becomes more

Table 2: Accuracy metrics for example city-wide SARIMA model fit

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

challenging and, as we will show soon, often requires collaboration between different companies.

Figure 3 (a) shows the relationship between the forecast accuracy and the number of rides reported within a district. Not surprisingly, we see that the accuracy is higher in districts with a higher number of reported rides. District-level predictions 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 a (negative) outlier in terms of accuracy in Fig. 3 (a). 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 pattern.

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 model 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

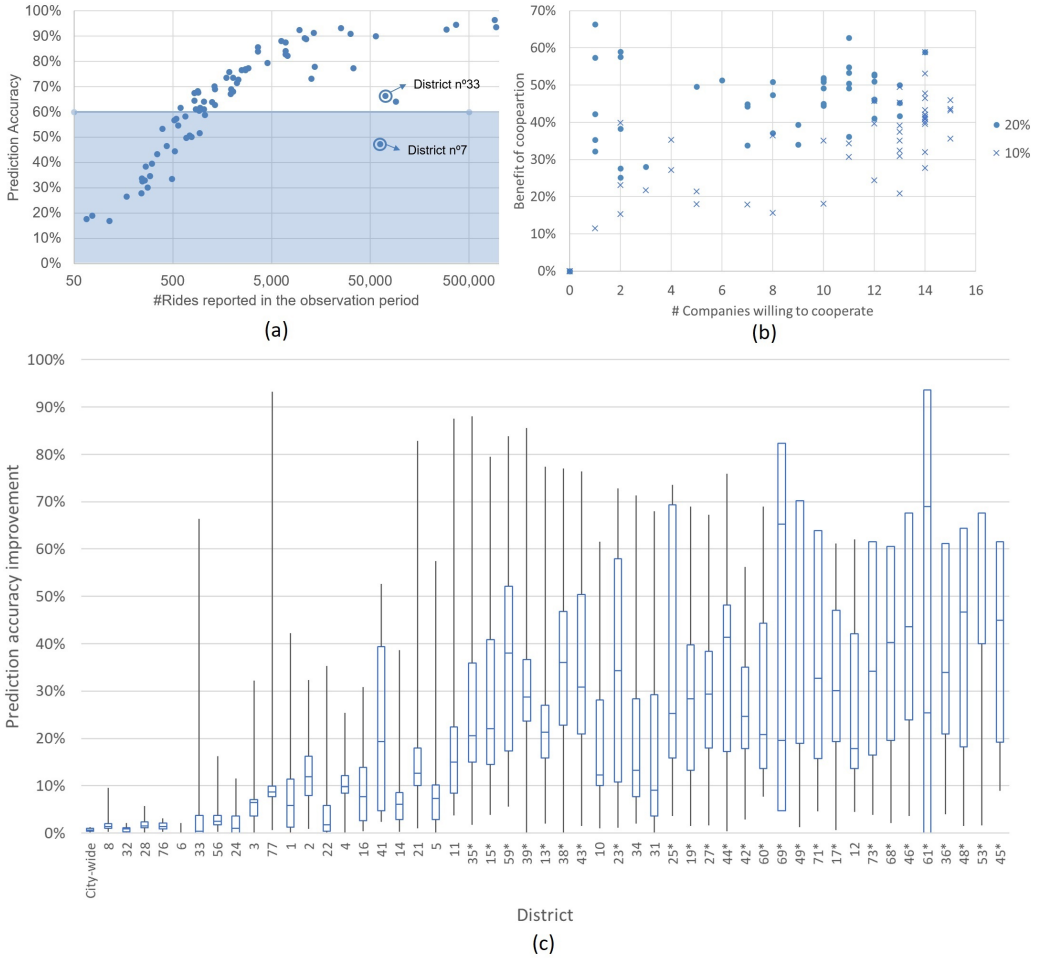

Figure 3: Demand prediction at district level. (a) Relationship between the level of accuracy achieved by district and the number of rides reported. Each point in the plot represents a district. (b) Benefit of cooperation *vs.*number of companies willing to cooperate in obtaining a better prediction for two different *cooperation thresholds* (20% or 10%) (c) Potential prediction accuracy improvement by cooperation at district level.

SARIMA model.[7]

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

In order to check if our findings at city level still hold at local level (particularly if every company is able to build their own accurate prediction model relying only on their data), we run the model in each of these 50 districts for all 16 companies. We compute for each one the *benefit of cooperation* as the difference between the accuracy of demand prediction using all companies and the average (across companies) prediction accuracy achieved by each of them on their own. For our analysis, we will assume that a company would be willing to cooperate if its prediction accuracy is improved by at least a minimum *cooperation threshold*. Figure 3 (b) plots for each of those 50 districts the average benefit of cooperation (Y-axis) vs. the number of companies willing to cooperate (X-axis), considering two different *cooperation thresholds*: 10% and 20%.

Looking deeper within district data, we find that in all the districts of Fig. 3 (b) 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. Also we see that in general, smaller companies tend to benefit more from the cooperation. It is also worth noticing that lowering the *cooperation threshold* leads to more companies willing to cooperate and a lower average *benefit of cooperation*.

Having taken a first look at the benefits for different companies in different districts, we turn our attention to those districts where inter-company collaboration makes more sense. Figure 3 (c) depicts box-plots (over companies) of the prediction 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. The plot shows clearly that in the most

[7]Areas like this may be amenable to a better prediction accuracy by more complex models using more 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.

popular districts (meaning that they report a large number of rides) the per-company benefits from collaboration are rather small as with city-wide results. 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 in order to achieve a higher demand forecast accuracy.

In summary, the plot shows that:

- there are 17 districts where most companies are able to provide accurate estimations (i.e., above 80% of the accuracy achieved using all the information) and, consequently where there are weak incentives for companies to cooperate.
- In 26 districts (marked with an asterisk in box plot (c) of Fig. 3) the average benefit of cooperation is at least 20%.
- In 33 districts the average benefit of cooperation is at least 10%

Focusing on the districts where collaboration makes most sense, we will now show how to compute the importance of the data that each company brings. We will do that via the notion of the Shapley that we will introduce next.

3.4. Introducing the Shapley Value

Establishing the individual player contribution to a collaborative game has long been a central problem of cooperative game theory. To this end, Shapley proposed that a players value should be proportional to their average marginal contribution to any coalition they may join [22].

Let $N = \{n_1, n_2, ..., n_{|N|}\}$ be a set of players and S_N be the aggregate data of N, with a value $v(S_N)$. The *Shapley value* is a uniquely determined vector of the form $(\phi_1(v), ..., \phi_{|N|}(v))$, where the element representing player n_i is given by

$$\phi_{n_i}(v) = \frac{1}{|N|!} \sum_{\pi \in \Pi} [v(S_{(\pi,i)}) - v(S_{(\pi,i)\setminus n_i})], \tag{2}$$

where π is a permutation representing the arrival order of the elements in set N, while $S_{(\pi,n_i)}$ represents the set of players that have arrived into the system before player n_i.

Unfortunately, the Shapley value has also been proven to be NP-hard for many domains [5]. Since it takes into account all possible coalitions, for each user, the number of terms scales with $2^{|N|}$, where $|N|$ represents the number of players, therefore it quickly becomes computationally infeasible.

In Ref. [19] the authors use Monte Carlo to approximate the Shapley value for computing the cost contribution of individual households to the peak hour traffic and costs of an Internet Service Provider (ISP). In that case, Monte Carlo had been used as a technique for approximating Shapley. Other recent works have presented approximation algorithms for Shapley for specific problems of lower complexity [6, 26]. Here we have tested both Monte Carlo and Truncated Monte Carlo methods, as well as Random Sampling and various Structured Sampling techniques (see appendix Appendix A.3 for more information).

3.4.1. A Toy Example

Consider a group of taxi companies agreeing to pool together the spatio-temporal data they have about demands for taxi rides in a city. One method to determine the value of a company is to observe how well the company is able to reconstruct the aggregate total, that is the data coming from all companies, by solely using its own data. 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 predicted signal and the actual aggregate signal, is measured. This error, or rather its opposite, the reconstruction accuracy, represents the value of the company (or coalition of companies). Aggregation leads to a highly non-trivial behavior of the value function, as defined above, and in the following, we will discuss a few particular cases in more detail.

Let us consider a toy example, depicted in Fig. 4a. A large number of companies combine their data, to produce a spatio-temporal 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, 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 complementary, de-

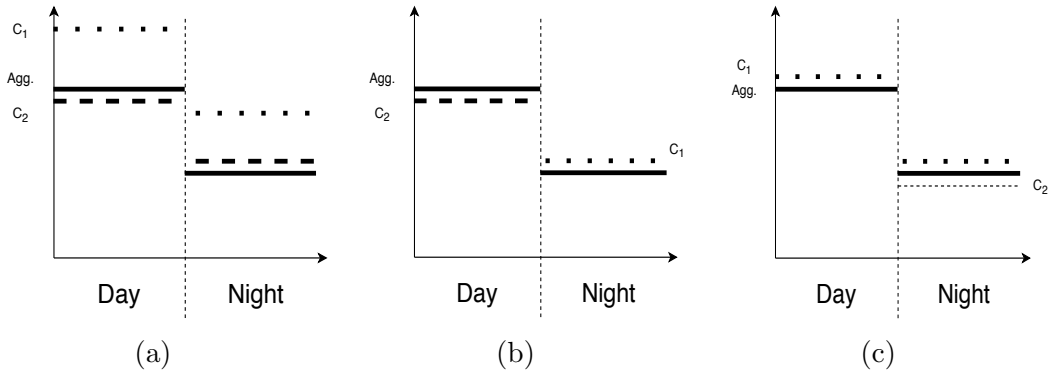

Figure 4: Data aggregation can influence its value in nontrivial ways. In figure (a) we have two companies, C_1 and C_2, active during en entire day. As C_2 has an average value closer to the reference aggregate (which represents data coming from many other companies), its data is better able to reconstruct the aggregate than C_1's data. In figure (b), the two companies act at different hours, and neither will be able to reconstruct the complete signal by itself, but by combining their data, they gain a significant advantage. In figure (c), the combination of data from a company active during the entire day with that of one only active during the night, will be detrimental to the task of predicting the complete aggregate, as C_2's data will distort the true activity gap between day and night. Not all data aggregations lead to added value.

picted in Fig. 4b. 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. Indeed, by combining their data, the resulting signal covers the entire time-span of the aggregate.

Data aggregation, however, does not always lead to an increase in value. Indeed, there are cases where adding a new provider to a coalition reduces the total value. A simple example is presented in Fig. 4c, one company, C_1 provides data spanning the entire day, and is also close to the total aggregate, while the other, C_2, only provides data during the night. By combining the two data sets, the overall predictive accuracy will drop because the absence of reports for day from C_2 will make most estimators believe that the traffic intensity gap between day and night is smaller than the real one.

The previous figure has shown through simple examples that depending on the particular characteristics of different datasets, mixing can be beneficial or not. In reality of course there are almost infinite ways in which different

spatio-temporal data may mix and that is why one cannot disregard all this complexity and substitute it with simplistic rules such as that the value of a company's data depends only on the number of the data points it provides without regard about what these data points look like and how they mix with the data points of others. *The advantage of using the Shapley value is that no such assumptions need to be made since the Shapley value considers by default all the possible ways in which datasets may be mixed.*

3.5. Computing the relative value of information at district level

For the 26 districts marked with an asterisk in Fig. 3 (b), taxi companies would benefit from an increase in prediction accuracy by combining their data. For each one of these districts we have computed the Shapley value of each company in the district. To do that we used the value function from Eq. 1, where the total aggregate is now the signal obtained by combining the data of all companies active in that particular district, and the predicted aggregate is the signal predicted by the SARIMA model, after being trained on the data from a particular coalition. For all such coalitions, the Shapley formula from Eq. 2 is applied.

Table 3 depicts the value assigned to each company by each one of these methods for the first 6 districts. Figure 5 shows the relationship between the number of rides and the Shapley value in our prediction at district level. Each point in such plot represents a company in one of the 26 districts. The Shapley value of a company in a district, represents the average marginal contribution to the accuracy of the predictor built by the data of a coalition of other companies for that district.

Observing Table 3 one may see that different taxi companies can have within the same district Shapley values that differ by several orders of magnitude. Also, the Shapley value of a given company may vary from district to district by a factor of more than $\times 10$ in some cases (see, for instance, companies 1 and 13 in districts 15, 17 and 19). Some companies have negative Shapley values in certain districts, meaning that they are bringing *on average* a negative contribution (*i.e.*, reduce the prediction 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 different companies just by the volume of data (rides) that

Table 3: Shapley value and number of rides (%) for a sample of districts

	13		15		17		19		23		25	
Co	SV	Rides	SV	Rides	SV	Rides	SV	Rides	SV	Rides	SV	Rides
1	11.8	11.3	11.2	2.5	14.0	8.3	2.0	3.4	11.2	3.9	14.1	8.3
2	1.8	1.2	1.8	0.8	0.0	0.5	1.5	0.5	3.5	2.9	4.5	2.3
3	1.8	0.3	1.0	0.3	0.2	0.5	0.3	0.0	0.4	0.1	0.3	0.0
4	0.0	0.9	0.4	0.2	0.2	0.0	0.4	0.1	0.7	0.3	0.1	0.1
5	1.4	1.2	2.3	0.9	0.4	0.5	0.7	0.8	3.0	2.4	0.5	0.9
6	20.0	49.2	16.4	37.9	28.0	56.2	24.1	38.6	21.0	42.9	21.8	56.1
7	2.4	1.0	1.1	0.4	0.2	0.4	0.2	0.5	0.9	0.1	1.1	2.4
8	1.5	1.0	1.1	0.8	0.3	1.4	1.5	0.5	3.2	2.2	-0.3	1.0
9	2.8	0.3	-0.3	0.2	0.0	0.2	-0.6	0.3	0.3	0.2	1.8	0.4
10	2.1	3.2	2.3	1.4	0.2	0.8	0.9	0.7	1.0	2.0	3.3	2.3
11	1.2	0.8	0.6	0.3	0.2	0.5	1.4	0.5	0.6	0.3	0.4	0.4
12	9.4	3.2	4.4	1.9	0.4	0.9	2.4	1.9	0.3	0.9	2.8	0.8
13	2.3	1.9	17.9	18.1	0.3	1.3	4.3	1.3	4.9	1.6	0.1	1.0
14	17.7	24.1	16.7	34.0	17.2	27.6	26.4	50.4	20.2	39.9	21.8	23.3
15	-0.4	0.2	0.4	0.1	0.8	0.3	0.4	0.1	0.0	0.1	1.0	0.3
16	0.8	0.3	0.2	0.2	0.0	0.8	2.4	0.5	0.6	0.2	-0.1	0.4

they contribute, we would be rewarding large companies, at the expense of smaller ones.

Later in section 4.5 we will address the calculation of payoff allocation in the case of individual drivers, and propose a way to manage positive and negative Shapley values to produce actual monetary payoffs.

3.6. 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 this cases, the data volume of a company does not reflect accurately its contribution to achieving a better forecast of future demand as given by its corresponding Shapley value.

4. Computing the importance of data in retail markets

In the previous section we developed methods for estimating the value of aggregate data held by taxi companies. In this section we will go a step further, and develop methods for estimating the value of data held by individual drivers. This will introduce additional challenges in terms of scalability of computation, since the Shapley value will now have to be computed over hundreds or thousands of drivers instead of few tenths of companies. Such

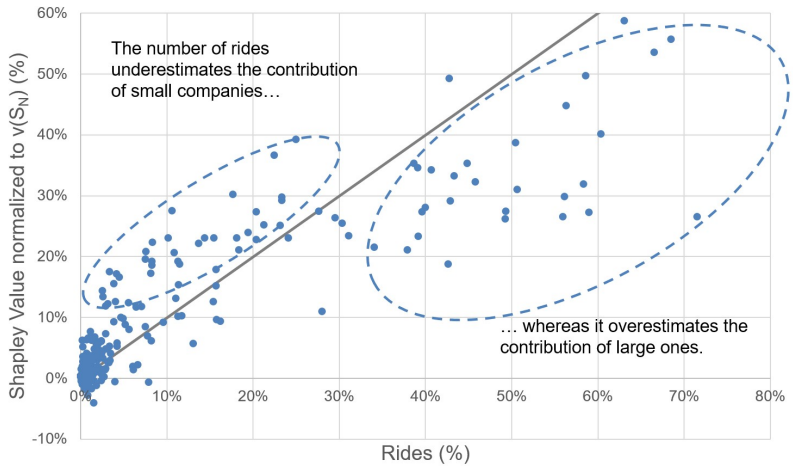

Figure 5: Shapley value vs. percentage of rides reported by company for a sample of districts. Each point represents a company in a district

a use case makes much sense since, in addition to market places for wholesale data such as IOTA or Airbloc, in the last few years we have also been observing the emergence of several specialised market places for individuals' data. Such marketplaces are often called Personal Information Management Systems (PIMS) in Europe and, in addition to allowing people to advertise and sell their data, they also offer other GDPR related functions such as data portability, erasure, and anonymity functions, among others. See Digi.me, MyDex, HAT, EarnieApp, Citizen.me, Meeco or OwnYourInfo for some examples of such PIMS.

4.1. Selling spatio-temporal data through a PIMS

To conduct our study of estimating the value of information held by individual drivers we will assume a simple model of a PIMS. The design space for PIMS is of course huge, but it is beyond the point of this paper to examine different alternatives. Hence in the rest of this section we will assume that the PIMS operates as follows:

- Drivers upload to the PIMS their rides each day.
- Buyers request from the PIMS to train their forecasting algorithm for spatio-temporal demand using data from real drivers.
- The PIMS uses a sufficient number of drivers' data to reach an accuracy threshold set by the customer.

- Buyers pay the PIMS.
- The PIMS keeps a small percentage of the payment and returns the remaining part to the drivers whose data was used in training the buyer's forecasting algorithm.

We will assume again that the PIMS uses a multiseasonal SARIMA prediction model as described in Sect. 2. To achieve the requested accuracy, the PIMS starts with a random number of drivers, trains the algorithm, and computes its accuracy over a test set. If the accuracy threshold is not reached, then the PIMS selects an additional set of drivers until it gets to the desired accuracy or fails to do so, in which case it informs the buyer that the request cannot be met.

In the subsequent sections we will first show how to compute the relative importance of each driver's data involved in the computation, captured by its Shapley value, and then how to transform it into an actual payoff for such data.

4.2. Computing the Shapley value for individual drivers

Computing the Shapley value directly requires evaluating the value of $O(2^{|N|})$ coalitions which may be feasible for few tenths of taxi companies but is clearly infeasible for hundreds or thousands of drivers in a city. For example, to do computation for the importance of the data of each driver in estimating the city wide taxi demand in Chicago we would need to consider 5000 drivers. Even in individual districts, if we were to consider the importance of each driver that has reported a ride we would still need to compute the Shapley value from hundreds and for some large districts, thousands of drivers.

To address the above scalability challenges we have implemented and evaluated a number of faster algorithms for approximating Shapley values. These include:

1. Truncated Monte Carlo approximation (TMC)
2. Random sampling (RS)
3. Structured sampling (SS), which plans the sampling upfront to ensure that all players appear r times in each position of the $r \cdot |N|$ sampled permutations of N.

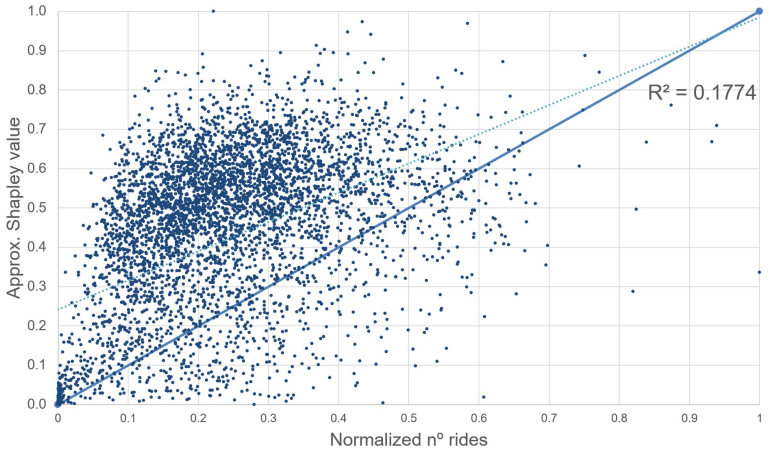

Figure 6: Approximate Shapley value vs. the number of rides across drivers

4. Truncated SS (TSS), which is a variant of SS that stops computing sample permutations once the accuracy reaches a certain threshold (*e.g.*95% of $v(S_N)$)

Having evaluated the above algorithms extensively (see details in appendix Appendix A) in terms of precision and robustness vs. computing time, we have selected the TSS algorithm since it achieved the best trade-off in all the datasets we tested.

4.3. City-wide results

We have computed a TSS Shapley value approximation for a set of $|N| = 4968$ taxi drivers that were providing service in Chicago during March and April 2019. We sampled $r = 8$ different permutations for each driver and applied a truncation threshold of 0.95. In this way we computed the contribution of each driver's data to the forecasting accuracy achieved by the multiseasonal SARIMA model in predicting the demand in the second half of April using taxi rides from the previous six weeks for training (T_o = Mar. 4th - Apr. 14th and T_c = Apr. 15th - 28th).

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. In fact, the linear correlation between both values is very low ($R^2 = 0.1774$).

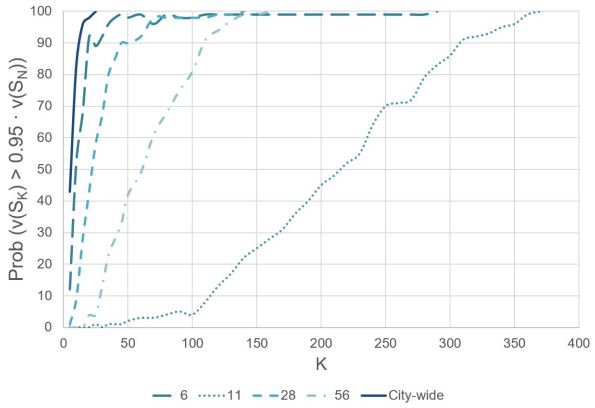

Figure 7: Probability of $v(S_K)$ exceeding $95\% \cdot v(S_N)$ vs the number of drivers in K

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.

4.4. District-level results

In the previous section we have shown that it is possible to build quite accurate demand forecasts only by using a very small number of drivers at city level. But what if someone, *e.g.*a customer of a PIMS as defined in Sect. 4.1 requires to build accurate demand forecasts at the district level?

To address this challenge we will first quantify the number of drivers that need to combine their data to get an accurate forecast of demand at district-level, and then proceed to estimate the value of each individual driver's data. Figure 7 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.

The plot show 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

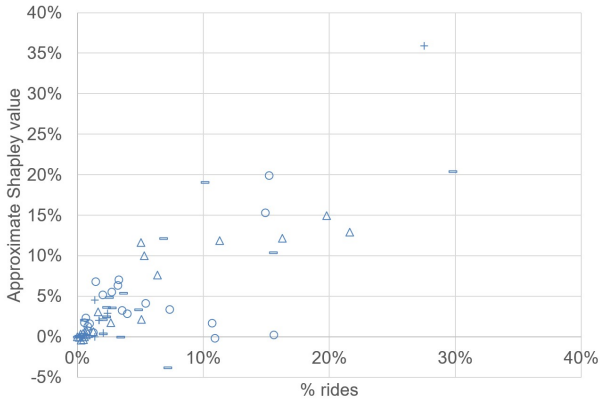

Figure 8: Scatter plot of approximated ϕ_i vs n rides (%) for tuples of 20 drivers in district 28 which yield more than $0.95 \cdot v(S_N)$ prediction accuracy.

known to be easier to forecast (see for example [16] in which the traffic of large backbone network links is easier to predict than the traffic of smaller access links). Achieving high forecasting accuracy in medium and small districts requires using the data from tenths 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 structured sampling approach discussed earlier in Sect. 4.2.

Nonetheless, we are able to compute the Shapley values for smaller sets of drivers whose data achieve an accuracy very close to $v(S_N)$ when combined. Figure 8 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 implied by the number of rides that he reports.

4.5. Translating Shapley values to actual payments

In the previous sections we analysed the relationship between a district size and the number of drivers that need to pool together their data in order to drive an accurate demand forecast at city and district level. Also, that the Shapley values assigned to individual drivers may vary significantly and that they cannot be approximated by using the number of rides that each one reports. In this section, we return to the PIMS model described in the

beginning of the section, and look at how we can produce actual monetary compensations on top of the Shapley values of individual drivers.

As noted in the beginning of the section, in order to fulfil a request from a customer, the PIMS selects random groups of drivers in batches of k until it arrives to a sufficient number of batches b that can achieve the desired forecasting accuracy demanded by the data buyer. Notice here that not all the drivers in the required minimum set of batches need to be considered for training the estimator of maximum accuracy. This may happen because some drivers may add noise instead of helping the estimator to be more accurate, as expected before in Sect.4.4. Therefore, out of the $b \cdot k$ drivers considered in b batches, the PIMS will identify the subset that achieves the maximum forecasting accuracy. Then payments can be assigned as follows:

- A fixed percentage of the overall payment will be kept by the PIMS for offering its service.
- The remaining will be split among individual drivers. This will be split into two parts:
    1. The biggest amount will go to the subset of drivers whose data is used in training the estimator of maximum accuracy out of the $b \cdot k$ considered drivers. Payments will be made in proportion to the Shapley value of each driver, neglecting those who deliver a negative ϕ_i. Notice that a negative Shapley value means that in spite of the fact that this driver's data is required in achieving the maximum accuracy among the $b \cdot k$ drivers, its average contribution in subsets of the drivers is negative.
    2. The remaining part will be split among all the users in the PIMS in equal amounts. The latter is done because the data of users not belonging to the $b \cdot k$ users is used in order to benchmark the accuracy of the forecasting obtained via the data of the $b \cdot k$ users, and thus they should also be entitled to a compensation.

The above is only an indicative scheme for computing actual payments from Shapley values. Coming up with the exact percentages paid to the PIMS, the users whose data is used for training, and the remaining ones whose data is used for bench-marking, involves several additional complications that go beyond our main task in the paper which is to validate the relative importance of different datasets. The exact percentages should depend on the prices asked (or offered) by data sellers (buyers), the market

power of a particular PIMS and its competition with other PIMS, with the size of the user-base of a PIMS (the smaller it is the highest the percentage to return to users in order to retain them and attract new ones), and others. We intend to study such matters as part of our future work with real users and data.

5. Taxi demand prediction in NYC

We have repeated the analysis using this time a dataset of taxi rides in New York City from Apr to May 2019 [8]. This dataset includes, for those three months, 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. Particularly in the case of NYC, more than 80% of taxi companies were able to predict demand with an accuracy of above 80% in 229 districts. Cooperation was identified to improve by more than 10% the accuracy of individual predictions for more than 75% of the companies in 27 of the smallest districts. We approximated Shapley values for those 27 districts for all companies. Finally, the model was not able to produce reasonable results in 4 districts due to a very reduced demand. In the areas where cooperation between companies made sense, the number of rides reported by each company was found again not to correlate well with the importance of the company as given by its Shapley value, as was also the case in Chicago.

Similar conclusions were obtained when analysing the value of individual drivers. Their Shapley value did not correlate well with the number of rides reported by each driver (R^2 ranging from 17% to 40%). 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. Related Works

The use of spatio-temporal data in transportation and smart city applications has attracted much attention from the research community. Works like [23, 27, 9] look at how knowledge extraction from spatio-temporal data can improve the effectiveness of transportation [24] and delivery services [25]. Despite, however, the large literature in the area, we are aware of only

[8] see `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`, last accessed January 2020

a single paper that studies data valuation matters around spatio-temporal data [3]. The focus of the paper, however, is different from ours since it is concerned with location-based advertising. Also, unlike our work, [3] uses *ad hoc* notions of value instead of the widely accepted Shapley value used in our work.

The research community is also actively working on matters of data marketplace design [1, 4] and pricing policies for data [17]. In parallel, many commercial PIMS (like Digi.me, MyDex, HAT, EarnieApp, Citizen.me, Meeco or OwnYourInfo) and data marketplaces (such as IOTA or Airbloc) are trying to open-up and commercialise individual and enterprise data. The idea of providing micropayments to users for their personal data has received a lot of public attention after the publication of "Who owns the future?" by Jaron Lanier in 2013 [14]. More recent work describes fundamental technological challenges that need to be addressed for the above vision to be fulfilled [15]. None of the above works has looked at valuation issues relating to spatio-temporal data.

A last body of related work has to do with applications, and computational aspects of the Shapley value. [11] has used the Shapley value to compute payments for providers of training data for different machine learning problems (not related to spatio-temporal demand prediction). Several works have looked at computational aspects of Shapley value and for efficient exact and approximation algorithms for particular types of problems such as recommendation, graph centrality, and others [13, 11, 6, 18].

7. Conclusions 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 future demand prediction for taxi rides in large metropolitan areas such as Chicago and New York. Our main result has been that the importance of each dataset cannot be deduced by just looking at the number of data points it includes but instead one needs to look deeper and consider the complex ways in which different dataset complement one another.

We have used the notion of Shapley value from coalitional game theory to compute the average marginal utility that a dataset is bringing when appended to a coalition of other datasets. This marginal utility is representing the improved forecasting that a dataset can achieve by complementing an

existing coalition of datasets (in our case past taxi rides). With the above tool at hand, we have analysed the values held by entire taxi companies aggregating data from all their drivers as well as by individual drivers that could sold such data to new marketplaces appearing for such purposes. Computing the Shapley value at the level of taxi companies can be done directly using the definition of what the Shapley value is. For individual drivers, however, we had to resort to structured sampling to approximate its value, since computing it exactly would be computationally infeasible.

As part of our future work we are looking at various related topics. First, how to design an actual marketplace for such data and how to compute actual prices and percentages paid to the platform and to the data providers in order to entice as many data providers and data buyers to use the platform. Such pricing problems are orthogonal to the relative importance of datasets studied in this paper but would benefit by having this information as a starting point for price setting and negotiation. Secondly, we are looking at value sharing in the context of more complex metrics than just the spatio-temporal footprint of demand. Such metrics include full origin-destination traffic matrices as well as congestion in road networks. Last, we are working on developing approximation algorithm for computing the Shapley value in spatio-temporal and other settings.

References

[1] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. 701726. https://doi.org/10.1145/3328526.3329589

[2] H. Akaike. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19, 6 (December 1974), 716723. https://doi.org/10.1109/TAC.1974.1100705

[3] Heba Aly, John Krumm, Gireeja Ranade, and Eric Horvitz. 2018. On the value of spatiotemporal information: principles and scenarios. Association for Computing Machinery, 179188. https://doi.org/10.1145/3274895.3274905

[4] Moshe Babaioff, Robert Kleinberg, and Renato Leme. 2012. Optimal Mechanisms for Selling Information. Proceedings of the ACM Confer-

ence on Electronic Commerce (04 2012). https://doi.org/10.1145/2229012.2229024

[5] Yoram Bachrach, Edith Elkind, Reshef Meir, Dmitrii Pasechnik, Michael Zuckerman, Joerg Rothe, and Jeffrey Rosenschein. 2009. The Cost of Stability in Coalitional Games. https://doi.org/10.1007/978-3-642-04645-2_12

[6] Sergio Cabello and Timothy M. Chan. 2018. Computing Shapley values in the plane. arXiv e-prints, Article arXiv:1804.03894 (Apr 2018), arXiv:cs.CG/1804.03894

[7] Javier Castro, Daniel Gomez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. Computers and Operations Research 36 (05 2009), 17261730. https://doi.org/10.1016/j.cor.2008.04.004

[8] Richard Durstenfeld. 1964. Algorithm 235: Random Permutation. Commun. ACM 7, 7 (July 1964), 420. https://doi.org/10.1145/364520.364540

[9] Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. MultiCell: Urban Population Modeling Based on Multiple Cellphone Networks. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2 (09 2018), 125. https://doi.org/10.1145/3264916

[10] Shaheen S. Fatima, MichaelWooldridge, and Nicholas R. Jennings. [n.d.]. A linear approximation method for the Shapley value. Artificial Intelligence ([n. d.]), 2008.

[11] Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. (04 2019).

[12] Mark T. Jacobson and Peter Matthews. 1996. Generating uniformly distributed random latin squares. Journal of Combinatorial Designs 4, 6 (1996), 405437. https://doi.org/10.1002/(SICI)1520-6610(1996)4:6<405::AID-JCD3>3.0.CO;2-J

[13] Ruoxi Jia, David Dao, Boxin Wang, Frances Hubis, Nezihe Gurel, Ce Zhang, Costas Spanos, and Dawn Song. 2019. Efficient task-specific

data valuation for nearest neighbor algorithms. Proceedings of the VLDB Endowment 12 (07 2019), 16101623. https://doi.org/10.14778/3342263.3342637

[14] Jaron Lanier. 2013. Who Owns the Future? SIMON and SCHUSTER.

[15] Nikolaos Laoutaris. 2019. Why Online Services Should Pay You for Your Data? The Arguments for a Human-Centric Data Economy. IEEE Internet Computing 23, 5 (12 2019), 2935.

[16] Nikolaos Laoutaris, Georgios Smaragdakis, Pablo Rodriguez, and Ravi Sundaram. 2009. Delay Tolerant Bulk Data Transfers on the Internet. In Proceedings of the Eleventh International Joint Conference on Measurement and Modeling of Computer Systems (SIGMETRICS09). Association for Computing Machinery, New York, NY, USA, 229238. https://doi.org/10.1145/1555349.1555376

[17] Sameer Mehta, Milind Dawande, Ganesh Janakiraman, and Vijay Mookerjee. 2019. How to Sell a Dataset?: Pricing Policies for Data Monetization. 679679. https://doi.org/10.1145/3328526.3329587

[18] Marius Paraschiv and Nikolaos Laoutaris. 2019. Valuating User Data in a Human-Centric Data Economy. arXiv e-prints, Article arXiv:1909.01137 (Aug 2019). arXiv:cs.SI/1909.01137

[19] Rade Stanojevic, Nikolaos Laoutaris, and Pablo Rodriguez. 2010. On Economic Heavy Hitters: Shapley Value Analysis of 95th-Percentile Pricing. In Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC 10). Association for Computing Machinery, New York, NY, USA, 7580. https://doi.org/10.1145/1879141.1879151

[20] Tjeerd van Campen, Herbert Hamers, Bart Husslage, and Roy 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), 3. https://doi.org/10.1007/s13278-017-0480-z

[21] E. J. Williams. 1949. Experimental Designs Balanced for the Estimation of Residual Effects of Treatments. Australian Journal of Scientific Research A Physical Sciences 2 (Jun 1949), 149. https://doi.org/10.1071/PH490149

[22] E. Winter. 2002. The Shapley value. Handbook Game Theory 3 (01 2002), 20272054.

[23] Tong Xia and Yong Li. 2019. Revealing Urban Dynamics by Learning Online and Offline Behaviours Together. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 1, Article Article 30 (March 2019), 25 pages. https://doi.org/10.1145/3314417

[24] Li Yan, Haiying Shen, Zhuozhao Li, Ankur Sarker, John A. Stankovic, Chenxi Qiu, Juanjuan Zhao, and Chengzhong Xu. 2018. Employing Opportunistic Charging for Electric Taxicabs to Reduce Idle Time. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 1, Article 47 (March 2018), 25 pages. https://doi.org/10.1145/3191779

[25] Yan Zhang, Yunhuai Liu, Genjian Li, Yi Ding, Ning Chen, Hao Zhang, Tian He, and Desheng Zhang. 2019. Route Prediction for Instant Delivery. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3, 3, Article Article 124 (Sept. 2019), 25 pages. https://doi.org/10.1145/3351282

[26] Kaifeng Zhao, Seyed Hanif Mahboobi, and Saeed Bagheri. 2018. Shapley Value Methods for Attribution Modeling in Online Advertising. (04 2018).

[27] Yi Zhao, Xu Wang, Jianbo Li, Desheng Zhang, and Zheng Yang. 2019. CellTrans: Private Car or Public Transportation? Infer Users Main Transportation Modes at Urban Scale with Cellular Data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3 (09 2019), 126. https://doi.org/10.1145/3351283

Appendix A. Testing Shapley Value Approximation Algorithms

The aim of this appendix is to introduce the algorithms that were analyzed and evaluated to select the most suitable approximation to the Shapley value. Since an exact calculation of the Shapley value requires an $O(2^{|N|})$ algorithm, the performance of the approximation algorithms to be used is critical. Having evaluated different candidate algorithms extensively, we have selected a structured sampling algorithm for it provides the best trade-off between accuracy and time, on all the datasets we have tested. Even though it is tailored to the behaviour of the value function defined for this specific problem, it will outperform naive methods in any problem where the marginal contribution of a player to a coalition strongly depends on its order of arrival.

Appendix A.1. Explaining the evaluation testbed

We computed the exact Shapley value for the daily prediction model, using a wholesale setting as described in section 3.1, in order to test the accuracy of the prediction algorithms, both at city level and in a medium-size district (particularly district 35). The following approximation algorithms were evaluated:

1. Monte Carlo (hereinafter, MC) approximation as stated in [11] evaluates the marginal contribution to coalitions extracted from a random sample of permutations of N until a convergence condition is met. We selected as such convergence condition a flag that controls whether the maximum relative variation of approximated ϕ_i is below an input threshold, which will range from 10% to 0.5%, before computing a new permutation.
2. Random Sampling (hereinafter, RS) as stated in [7] using a different number of $r \cdot |N|$ sample permutations, where r will range from $r = 1$ to $r = |N|^2$.
3. Structured Sampling (hereinafter, SS), tailored to problems where the position of a *player* in a coalition strongly determines their marginal contribution, based on [10, 20]. SS ensures that all companies appear r times in the first positions for a set of $r \cdot |N|$ permutations.

Given the stochastic nature of Shapley approximation algorithms, we tested each one 50 times for each set of input parameters and obtained the approximate Shapley value for the 16 companies. We compared the performance of the aforementioned algorithms in terms of:

- Accuracy, measured as the average average[9] absolute error (hereinafter, AAAE) and average average percentage error (hereinafter, AAPE) compared to the exact Shapley Value by company.

- Robustness, measured as the average average[10] standard deviation (hereinafter, AASTDE) of the outputs of an algorithm and a certain set of parameters.

- Time to execute (TtE), measured in terms of the number of training-prediction cycles computed.

The convergence threshold in the case of MC and r in the case of RS and SS allow one to define the sample depth and affect both to the execution time and to the accuracy of the approximation.

In all cases, we tested non-truncated and truncated versions of the MC, RS and SS algorithms (we will refer to them as truncated-XXX algorithms, or in short-form TMC, TRS and TSS). Truncation of execution above a certain *truncation threshold* $(v(S_N) - \epsilon)$ works in the following way: while evaluating a permutation π of set N, if it holds that for the coalition of the first j players, $\pi_j \subseteq \pi, \pi_j = \{\pi[1], ..., \pi[j]\}, v(\pi_j) > v(S_N) - \epsilon$, the rest of members $k \in \pi - \pi_j$ are considered to bring a zero marginal contribution. Truncation helps the algorithm reduce the time it takes to execute but also decreases the accuracy of the approximation.

Appendix A.2. Explaining structured sampling (SS) algorithm

As regards to *"structured sampling"* (SS) approximation to Shapley value, the **Algorithm 1** box provides a detailed description of the algorithm. Unlike random sampling, it plans the sample permutations upfront so as to ensure that each player $i \in N$ appears r times in each position of the sampled permutations. This reduces the randomness of the sampling process and increases the performance especially when the marginal contribution of a player i in a permutation π is significantly determined by its position.

We resort to Latin squares in the sample process. A Latin square LS of order $|N|$ is an $|N| \cdot |N|$ array with elements of a set N, in such a way that

[9]First average error across companies for each test, then average the average error across all executions.

[10]First average standard deviation of the approximate Shapley value across all executions, then average the average standard deviation across companies.

Algorithm 1 Structured sampling approximation algorithm

inputs: $y(t)$ train data for each —N— sources in the set N, accuracy test procedure v, rounds of permutations to evaluate r and truncation threshold ϵ
Initialize Shapley value vector $\phi_i = 0 \forall i \in N$
Initialize the set of sample permutations $P = \emptyset$
Create a $|N|x|N|$ Latin square - LS
$Q \leftarrow N$
for all $i \in \{1...r\}$ **do**
  $Q \leftarrow \text{shuffle}(Q)$
  $P \leftarrow$ set of $|N|$ permutations of Q according to the order defined by LS
end for
$t \leftarrow 0$
for all $\pi^t \in P$ **do**
  $t \leftarrow t + 1$
  $v_{j-1} \leftarrow 0$
  for all $j \in \{1...|N|\}$ **do**
    if $v^t_{j-1} <= v(S_N) - \epsilon$ **then**
      $v^t_j \leftarrow v(\pi^t_j)$
    else
      $v^t_j \leftarrow v^t_{j-1}$
    end if
    $\phi^t_{\pi^t[j]} \leftarrow \frac{t-1}{t} \cdot \phi^{t-1}_{\pi^t[j]} + \frac{1}{t} \cdot (v^t_j - v^t_{j-1})$
  end for
end for
outputs: Approximation to the Shapley value of each data source i: $\phi_1 ... \phi_{|N|}$

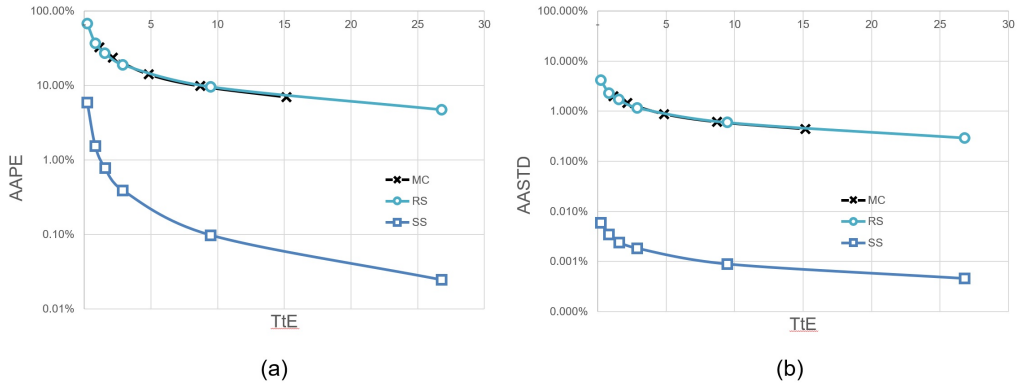

Figure A.9: Testing of Shapley value approximation algorithms for demand prediction models at city level (—N—=16). (a) Accuracy (AAPE) *vs.*execution time. (b) robustness *vs.*execution time

each element n_i occurs precisely once in each row and column of the array [12]. Latin squares have been extensively used in experiment planning [21]. As shown in the **Algorithm 1** box, we shuffle the elements of a random permutation Q of N according to the order defined by such Latin square to produce r different sets of $|N|$ permutations with the aforementioned properties.

The referenced shuffle algorithm is the modern version of the FisherYates shuffle, designed for computer use by Richard Durstenfeld [8]. Such algorithm runs in $O(n)$ time and is proven to be a perfect shuffle, assuming a reasonably good random number generator.

Appendix A.3. Evaluating Shapley value approximation algorithms

Fig. A.9 shows a comparison of MC, RS and IS in terms of accuracy and robustness. In subplot (a) we depict the AAPE as a function of the number of sub-coalitions evaluated, which determines execution time. Subplot (b) shows the AASTD as a function of the execution time. Please note that Y-axis is logarithmic in both subplots.

In all cases the more combinations are evaluated, the more accurate and, especially, more robust the results are. Nonetheless, it is clearly shown in this case that the SS outperforms both RS and MC, meaning that the planning of the sample permutations delivers a consistent output across executions which is also closer to the exact Shapley values.

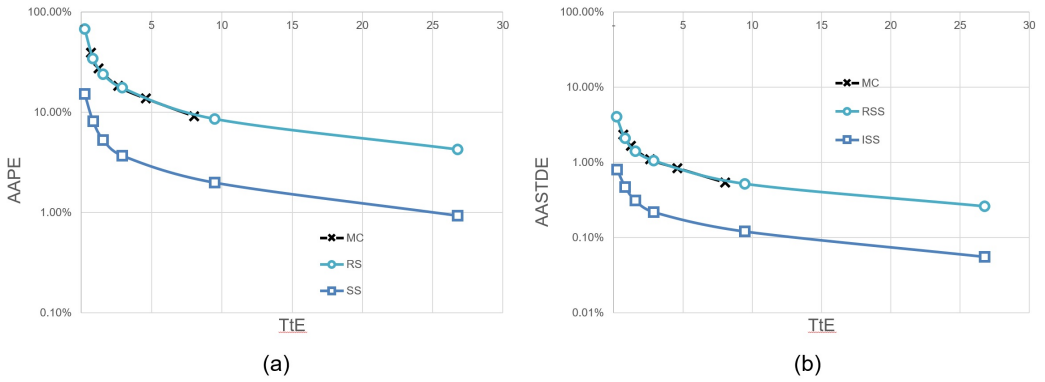

Figure A.10: Shapley value approximation algorithm testing for demand prediction models for district 35 in Chicago and N=16 companies. (a) Accuracy (AAPE) *vs.*execution time. (b) robustness *vs.*execution time.

Since city wide demand prediction considering data from companies is a very special case [11], we ran the same test using the inputs of a medium-size district, which is yielding very different Shapley values for each company, to prove whether or not the previous conclusions hold, in a scenario where the standard deviation of ϕ_i across companies is relevant. Figure A.10 shows the results of this analysis.

As expected, the difference between SS and the naive versions of MC and RS in terms of robustness and accuracy decreases in cases where ϕ_i values are very different. MC takes more time to converge and both RS and SS results show higher AAPE if compared with the first most favourable case. However, the SS algorithm showed to clearly outperform both MC and RS also in this situation.

In the light of the results obtained, we have selected SS as the best algorithm, since it is able to to approximate the Shapley value with a 10% average error in $O(|N|^2)$. This we consider sufficient for the purpose of a value-based payoff distribution. In case finer accuracy is required, SS is able to estimate the Shapley value with a 4% of error in $O(|N|^3)$.

[11] If we recall section 3.2, all companies have enough data to independently predict the shape and average mean of the aggregate demand, with at most a 10% maximum average error, meaning that the company that appears in the first place in the permutation is bringing all the value. This might be a best case for SS.

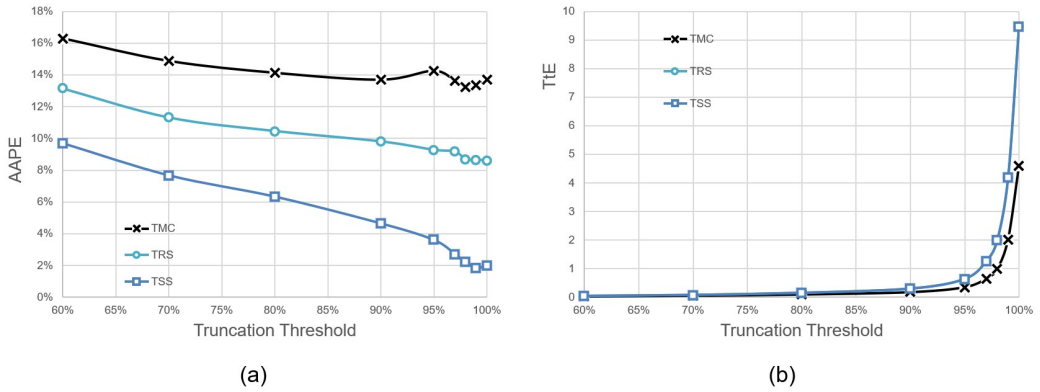

Figure A.11: accuracy and TtE of ϕ_i vs. truncation threshold in approximation algorithms. In this case, SS and RS overlap in the (b) plot.

Appendix A.4. Evaluating the impact of truncation on accuracy

For computing the Shapley value for a large number —N— of *players*, and given the specific behaviour of value in demand prediction problems, truncation proves to be an important feature to speed up the execution without necessarily distorting the output of the algorithm. In fact, if according to section 4 by taking only a small percentage of all drivers we are able to quite accurately predict demand in most cases, then why spend our valuable computing time evaluating the marginal contributions of additional *players* once our prediction has reached a 95% of the maximum accuracy?

We have computed approximations to Shapley value for the following truncation thresholds: 0.6, 0.7, 0.8, 0.9, 0.95, 0.97, 0.98, 0.99 and 1. We have run TMC, TRS and TSS in the same district as we did in section Appendix A.3 for $|N| = 16$ companies, 50 times for each ϵ and using a convergence threshold of $0.01 \cdot v(S_N)$ for TMC and $r = 64$ for TRS and TSS.

Figure A.11 shows the effect of truncation in both accuracy (a) and execution time (b) for each of the three algorithms. According to our results, SS is significantly more sensitive to truncation but by tuning r and ϵ it is possible to control the trade-off between accuracy and execution time. We chose to use a *truncation threshold* of $0.95 \cdot v(S_N)$ since it divides the overall execution time by 16 while it only duplicates the percentage error.