Data-driven anonymization process applied to time series

Vincent Thouvenot Thales Communications & Security
Damien Nogues Thales Digital Factory
Catherine Gouttas Thales Digital Factory
firstname.lastname@thalesgroup.com

Abstract

Digital transformation and Big Data allow the use of highly valuable data. However, these data can be individual or sensitive, and represent an obvious threat for privacy. Anonymization, which achieves a trade-off between data protection and data utility, can be used in this context. There is not global anonymization technique which fits at all applications. Here, we describe a data-driven anonymization process and apply it on simulated electrical load data.

1 Introduction

The following paper is mainly written for a task of dissemination about anonymization and good pratice about it. Indeed, if anonymization is quiet well known from academic point of view, it is not still the case from France/Europe's industrials' one. However, privacy protection is a fundamental growing task for them. Digital transformation brings creation of global datalakes and allows development of new valuable business. Moreover, some Governments force an increasing putting in Open Data, which should promote the opening digital knowledge and ensure an open valuable numeric ecosystem. At the same time, European Union sets up rules to protect citizens, which establish that citizens have protection right for their individuals data. The data have to be fairly processed for specific purpose, and with individuals' agreement. It is an important point because keeping a maximum of personnal data for future non specified mining task which should appear through future methods is inconsistent with this previous rule. People have right to access and rectified their individual data. With 2016 regulation, applied from 2018, any company offering goods or services (including cloud services) to EU citizen may be subject to regulation. Besides this legal context, Big Data technologies enables the treatment of massive, dynamic and unstructured data, and facilitates data crossing, weakening privacy protection. The data concerned can be personal (name, address, etc), and allow to (almost) directly identify an individual. Sensitive data, like religious or political beliefs, pose a risk for individual privacy too. Smartphone, smart object, loyalty card, online purchase, social media : there are large sources of individual and sensitive data, which lead to an obvious risk for privacy. Protect privacy means avoiding the isolation of an individual, the correlation of some information from different datasets for one individual and the possiblity to obtain information on an individual through exogeneous variables. Despite appearances, the trade-off behind anonymization is not an easy task. In datasets, it can have some identifiers. Just delete or encrypt them is generally not efficient (see e.g. Hansell (2006); Narayanan and Shmatikov (2008)). Others information can be quasi-identifiers which allow re-identification when they are crossing (e.g. in a health dataset, sex and age). There are sensitive data (e.g. disease in health context). Finally there are some remaining parameters. In addition, the trade-off depends of three parameters. First it depends of data typology. Bank transaction data (e.g. see Ezhilarasi and Hariharan (2015)) will not require the same treatment that unstructured data like social media data, which ask to hide metadata, the identifying content and the relational graph data (e.g. see Zhou et al. (2008); Chester et al. (2013)). Second, the trade-off depends of the future use of data. Third, the anonymization strength is time dependent. Indeed, new datasets and new re-identification methods can be used to attack privacy. So, an admissible anonymization methodology can be unadmissible few months/years after.

We illustrate an anonymization workflow on (simulated) electrical smart meter data furniture, which are a symbolic example of sensitive and individual data whose exloitation possibility is new and illustrates new needs of anonymization. In France, smart meters, which are currently deployed and whose deployment ended in 2020 will allow to gather infra-daily household electrical load. These data are by nature personal and sensitive. Infra-daily individual loads allow to detect if and when someone is at home and can increase risk of burglary for instance. The individual habits (see e.g. (Blazakis et al., 2016)) can be detected and so sensitive data like religion (e.g. during Ramadan) or some illegal activities (e.g. very particular load pattern with cannabis plant) derived. Provide these data is a complicated challenge. Indeed, household electrical load will be available at different level of time granularity (e.g. see Tudor et al. (2013); Buchmann et al. (2013)). To simplify the context, in France, infra-daily load will be available to the individuals, which can choose to temporarily give access to these data to a third party. Electric distributor will gather daily data (even infra-daily in particular context). Provider will access to monthly data. Althougt infra-daily data are noisly, having access to identified daily (even monthly) data allow easily to re-identify infra-daily consomption (e.g. see (Tudor et al., 2013)). In this context, just hide direct identifyer will be inadequate. On the other side, these data have a strong valuable potential and many actors are interested by them, for instance, distributors to manage their network and achieve maintenance task; local communities to improve their urban policy; providers to propose new more adaptive pricing; or start-up to popose individuals some services to optimize the consomption. Based on the future use of data, it is not necessary to keep the same information. For instance new adaptive pricing and dimension the networks through household electrical load need different information. Finally, the currently acceptable methods can be questionned when data from gas smart meters, which are currently deploying too, will be available too. Indeed, gas meter data depends of similar phenomenon that electrical meter data.

The article is divided in four sections. In Section 2, we give a short survey about anonymization. In Section 3, we describe our global anonymization process, from data gathering to dataset publication. Section 4 presents a simulator which are respectful of French electrical smart meters anonymization task. We use simulate data because we have not access to real data. We applied the Section 3 process in Section 5 on Section 4 simulation.

2 A short survey about anonymization

For a dissemination task, it seems important to have a brief discussion about the differences between anonymization and encryption because confuse the two is a common mistake. Data encryption consists in using some mathematical algorithms to transform data. The process can be reversed with the good algorithm and the encryption key. It could be used to transfer data between two entities. The encryted data are still individuals and so still personal data if original data are personal. Although encrytion can be useful to be one of the components of de-identification, it is neither necessary nor sufficient for doing anonymization.

Pseudonymisation, which consists of hiding identifying metadata can be not efficient (see Danezis (2013)). De-identification falls into two categories of techniques: transform data to have unreal individual, and aggregate and generalize individual, where data provided symbolizing an individuals set. Techniques can be used and combined.

Permutation techniques and puzzling approaches deconstruct, transform and/or change the data design (see e.g. Agrawal and Srikant (2000), Zhang et al. (2007)). Noise addition techniques are popular. For instance, Dufaux and Ebrahimi (2006) randomly transforms video representation by inverting some signs in decomposition coefficients and applies it to privacy protection in video surveillance. Aggarwal and Yu (2008) proposes a survey about randomization methods. Classical randomization has some advantages. Noise is independent of data and does not require entire dataset. It can be applied during data collection and on distributed system. Liu et al. (2008) proposes an survey of attacks techniques on privacy obtained by perturbations methods. When the anonymizer adds an additive noise, the attackers can use methods like (spectral, singular value decomposition, principal component analysis, etc.) filtering, maximum a posteriori estimation, or distribution analysis. Under good conditions, multiplicative pertuba-

tions have good properties, for instance preserve Euclidian distance. An attacker who knows a sample of input and output or has some independent samples from the original distribution can reverse the anonymization. Differential privacy (see Dwork (2006), Dwork and Roth (2014)) is a very popular form of noise addition. Here, we add a random noise in such a way it makes a mechanism which produces the same output with almost similar probabilities when we consider two adjacent inputs. The basic process to achieve differential privacy is to sampling without replacement the dataset and adding fictive individuals. Differential privacy allows to work on privacy loss and bound the risk. Chatzikokolakis et al. (2013) applies differential privacy on mobility trace and tries to develop a mechanism more efficient than just adding independent noise. McSherry and Mironov (2009) proposes a framework of differential privacy to produce recommendations from collective user behavior in Netflix Prize dataset.

Privacy can be protected by creating individuals sets. K-anonymization (see Sweeney (2002)) consists of generalizing quasi-identifying information to force having at least k individuals with the same values. K-anonymity can be broken when all the individuals of (at least) one class have the same sensitive data. L-diversity (see Machanavajjhala et al. (2006)) forces each class to have at least l different values of the sensitive data. In t-closeness (see Li et al. (2007)) the sensitive data in each class has to respect its distribution in the total population. Generalization and suppression is NP-hard. Moreover, as expressed in Domingo-Ferrer and Torra (2005), generalization and suppression can be not adapted for ordinal categorical and for continuous attributes. Domingo-Ferrer and Torra (2005) proposes to use microaggregation for this task. In microaggregation data are partitioned into several clusters of length at least k with similar records. Then, we apply an aggregator operator (e.g. mean or median for continuous variable) to compute the centroid of each cluster. Besides clustering method, microaggregation has two important parameters: the minimum dimension of each cluster, adjusts the level of privacy protection and the function allowing computation of aggregate value, which is linked with future data utility and protection level. The aggregation function can be mean, sum, median, quantile, partial autocorrelation function, time slicing profile, density, etc. Domingo-Ferrer and Torra (2005) partitiones data throught Maximum Distance to Average Vector (MDAV) algorithm. Aggarwal et al. (2006) proposes microaggregation where some atypical individuals can be not clustered and so not published. Byun et al. (2007), Lin and Wei (2008), Li et al. (2002), Xu and Numao (2015) and Loukides and Shao (2008) proposes greedy heuristic to achieve k-anonymity through clustering with not NP-hard complexity. Bergeat et al. (2014) compares two software allowing k-anonymization on a French health dataset of more than 20 million records. Gedik and Liu (2008) uses k-anonymity to protect mobile location privacy.

When working on time series, previous techniques can be used. For instance, Shou et al. (2013) proposes what they named (k, P)-anonymity to preserve pattern in time series. In electrical household load protection, Chin et al. (2016) proposes to solve an optimization problem with two components : first is about information leakage rate of consomer load given grid load and second is about the cost of errors. Shi et al. (2011) proposes a differential privacy form to protect time series. Zhang et al. (2015) proposes noise generation to protection cloud data. Hong et al. (2013) proposes a survey about time series privacy protection.

After de-identification, it is important to measure de-identification degree. Venkatasubramanian (2008) surveys the metric proposed to measure privacy and privacy loss. Authors divide measuring privacy into three categories: statistical methods, taking account variance of perturbated variable, probabilistic methods, considering information theory and Bayesian analysis, and computational methods, coming from the idea of a resource-bounded attacker and measuring privacy loss in function of the information available for such attacker. Tóth et al. (2004) works on message transmission and analyzes two entropy based anonymity measures. Authors measure separate global anonymity, which quantify the necessary effort to fully compromise dataset, that we name latter journalist scenario, and local anonymity, which quantify the probability that transmission of one user are compromised, that we name prosecutor scenario. Gambs et al. (2014) works on attacks on geolocated data. Nin and Torra (2009) proposes a framework of protection and re-identification for

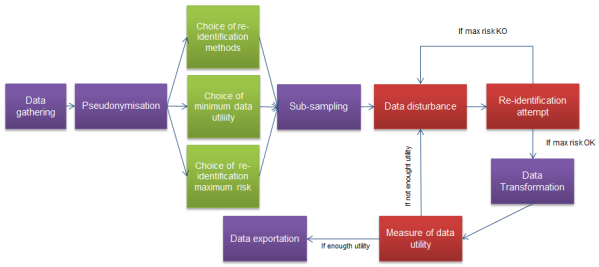

Figure 1: Global anonymization process

time series. Ma and Yau (2015) proposes some information measures for quantifying privacy protection of tme-series Data.

Interested reader can find in the first-rate book Torra (2017) the stakes of data privacy and techniques associated.

3 Anonymization process

Each anonymization task has a specific and a generic part, and so is unique. In this section we describe our global data-driven anonymization process (see Figure 1) which allows separating the two parts of the process.

After data gathering we have to tag data in function of theirs categories : identifier, quasi-indentifier, sensitive data, and remainded parameters. Next step consists in data pseudonymisation. Identifiers have to be hidden (deletion, encryption, etc.). Obviously, that is generally not enough to insure privacy protection. We have to establish one metric to measure data protection and another one for data utility. As anonymization could not be total and perfect, we have to choose the threshold of re-identification acceptance. Despite de-identification process there are still residuals risk, which has to be compared to the benefits.

It is necessary to build a re-identification framework, which is driven by the context and has to be realistic. That means the worst case situation, where an attacker is almost all-knowing, is probably not realistic and decreases the efficiency of the trade-off utility data / privacy protection. The re-identification framework depends of many parameters. The attacker type must be determined. Its resources will depend of who he is (e.g. a member of the organization which anonymizes data and so has access at plenty data to attack anonymized dataset, a member of a near organization which has access to similar data which can be crossed, a Machine Learning expert which can deploy efficient re-identification models, a neighbor which has access to contextual data, etc.). Many reasons can motivated the attacker (retrieve information about individual to do aggressive commercial supply or burglary, harm the organization which manages the anonymization to recover data governance for instance, show its capacity in re-identification, etc.). The re-identification framework will depend of the attack category. Crossing anonymized data with dataset which is not anonymized is a classical way to try re-identification. Information in anonymized data can be used too (e.g. anonymized Internet requests can be identified by crossing location and interest of individual). It will depend of the chosen re-identification meaning. It can mean identify an individual or identify sensitive information about an individual. For instance, if household load have been anonymized by pseudonymization then adding a noise, the noisy curve can be identify to a customer. If we do microaggregation, it could be possible to find the customer cluster and possibly deduce probable sensitive information and behavior for the customer. The re-identification framework will depend of the technique to measure re-identification risk. To be valuable, attackers have to be confident in their re-identification. When we compute the risk of individual identification, true positive are not the key performance indicator. Here, true positive means the good individual from anonymized data have been identified at the individual from not anonymized data. However, this individual from not anonymized data can be identified at many other individuals in anonymized data. That decreases re-identification risk. Of course, identification errors decrease the confidence too. The risk have to combine all these information. Lastly, re-identification framework will depend of the re-identification scenario. Many are possible : we can target all or almost all individuals when we know they are in the dataset (journalist scenario), we can target one or some individuals (prosecutor scenario), we can try to distinguish studies with and without one individual, etc.

Then, as anonymization is a trade-off between utility and protection, we have to choose the minimum utility of data. We have two case. Data could be provided to a third party to answer to a specific need. Only necessary information, and not more, have to be provided. After anonymization, the study has to be possible. For instance, if a elec-

tric provider want to do new daily pricing, they will only need precise daily profile. Data could be given without specific need, for instance to push data in Open Data. We need to compute a metric of privacy cost. For instance, when we add a noise in time series, we can compute a signal to noise ratio.

Lastly, as anonymization can not be perfect, we have to choose the limit of acceptance for re-identification risk. It determines the trade-off achievement point. It depends of the level of individuality and sensitivity. The choice is driven by the exportation model used at the end. We can choose a Publish and forget model, where typically data are provided in Open Data. Then, it is (almost) impossible to stop data sharing. Another model is Data Use Agreement model where agreement decides what the third party can do. Finally we can use a closed model where data are in a closed environment and the third party has only access to the results of its requests.

To avoid scalability problem during non industrial step we can do an optional sub-sampling. De-identification methods, which depends of data typology and future data use, are applied. Then, we measure the re-identification risk. When the risk is lower than the authorized maximum, we transform the entire dataset. Then, we measure the utility of anonymized data. If the minimum utility is not respected, we re-start all the steps of this paragraph, else we export data in the chosen model.

4 Appliance context : simulated electrical load

Electrical household load simulation has a rich literature. Many authors use variant of Markov Chain (e.g. Labeeuw and Deconinck (2013), Muratori et al. (2013), McLoughlin et al. (2010)). McQueen et al. (2004) uses Monte Carlo simulation model of load demand taking into account the statistical spread of demand in each half hour using data sampled from a gamma distribution. Paatero and Lund (2005) uses bottom-up load model for generating household load profiles. Pompey et al. (2015) trains Additive Model (see Hastie and Tibshirani (1990)) to achieve massive-scale simulation of electrical load in Smart Grids. Additive Models have yet proven their efficiency to model and forecast electricity load at aggregate level (see Pierrot and Goude (2011) in France, Fan and Hyndman (2012) in Australia) as at local level (see e.g. Nedellec et al. (2014)).

We simulate three curves types : we name the first “second house load”, which are almost constant with a random noise added, the second “little professional”, which are represented by segment curves with a load almost null the week-end and the night and almost constant during the day, where the jump intensity and activities period are randomly chosen, and the third “household load”, with calendar and thermic components. To simulate the last, we apply similar idea that Pompey et al. (2015). We train simulation models on GEFCom 2012 dataset (see Hong et al. (2014)). The dataset comes from a Kaggle challenge and contains the hourly load demand of 20 local areas in USA and the temperature of 11 weather stations. We train many different Additive Models on this dataset whose features sets contain calendar parameters (type of day, number of day since the beginning of the year, etc.) and a random number of raw and smoothing temperatures. They are trained on different period. Then, we compute two months of their forecasting with random translation of features (e.g. one additive model design is translated of one hour and its temperatures of two Fahrenheit), to introduce variety in simulated load. We train some quantile additive models too (see Gaillard et al. (2016)) to simulate some extreme behaviors. Our simulated data are smoother than real individual load but it is not a real problem here. Indeed, for our appliance, it is important that simulation respects some assumptions. We assume there are different levels of load and three curves families and the consomptions are almost uniqueness even at low frequencie as shown in Tudor et al. (2013). To respect an uniqueness assumption, we arbitrarily impose that daily aggregation during one week of two curves of “household load" rounded to the thousands can not be equal. We choose the daily time scale of aggregation because we are faced with an attacker using daily data and by considering one week, we integrate the weekly cycle which are important when considering electrical information.

We assume that a provider needs to refine a pricing for an area of its customers. It needs some infra-daily information (peak of demand, profile, etc.). In our scenario, a potential attacker has access to identified daily data. We assume the exportation model is a “publish and forget model”, which explains the possibility for an attacker to

Method	Parameter	Utility
Random noise	Signal to noise ratio, noise familly	trend, anomaly detection, total scope
Permutation	Type and quantity of permutation	total scope
Transformation	type of transformation (standardization, etc.)	total scope, trend, anomaly detection
Time slicing	time scale	total scope, anomaly detection
Differential privacy	sampling probability and fictive curve	total scope, trend, anomaly detection
Microaggregation	clustering, clusters dimension, aggregation function	total scope, clusters trend, profile
Scope aggregation	scope, aggregation function	total scope, scope trend, profile

Table 1: De-identification for time series

have daily data access.

5 Appliance of anonymization process

After data gathering, we have to tag the data. Here we assume data have two attributes: an individual identifier and simulated time series. First, we pseudomyze the individual identifier by substituting it with random numbers without replacement to ensure uniqueness. We have only one quasi-identifier, which is the sensitive data too, the simulated time series. In our simulation framework, the highest level of attackers can be a competing organization which has identified data at a fine granularity level (daily data), with good level of optimization and computing whose objective is to find information about customers' behaviors to make aggressive supply. Data are provided to answer to a specific need of a third party: having data to etablish new pricing. For this topic, microaggregation is relevent. Then, we protect privacy throught the minimum dimension of clusters. The components of the trade-off are the choice of clustering methodology, the choice of the dimension of clusters and the choice of the aggregator operator. With microaggregation, an attacker could, at worst, identify probable customers behaviors. More the minimum dimension of clusters is weak or more one load participates at the building of a cluster, more the attacker can be certain of its deductions.

5.1 Statistical and Machine Learning Setup

Table 1 presents some techniques which can be used to protect time series. Permutation, which consists to exchange data from one curve to another, is a form of noise introduction and can create unrealistic curves. Transformation can be smoothing, standardization, etc. whose objective is to erase some individualities. Time slicing breaks individual trends. Differential privacy can be completed by post-treatment to improve protection. The re-identification risk of the first five techniques of Table 1 are about customer identification and can be studied with classical time series identification and classification techniques (Deep Learning, Ensemble Method, K Nearest Neighbors, etc.). Moreover, some methods are reversible. For instance, denoising methods can be applied for the first one. For deterministic transformation, the perturbation can be inverted with the knowledge of transformation parameters. Attempting to re-build chronological can reverse time slicing. Sensitive information can be refund from the two last methods when attackers can identify the cluster of a customer. Here we work to provide data to a provider who wants to make a new pricing. It needs precise profile and we choose to use microaggregation. As a provider will not propose one tariff by individual, but many tariffs depending of large group, we do not need precise individual data.

5.2 Chosen methodology

We use time series clustering (see Liao (2005), Rani and Sikka (2012)). As explain in these surveys, there are many ways to cluster time series. First, clustering algorithm can directly be applied on raw data. However, it can be inefficient because of noisy data. Second consists to extract features from time series and applies clustering algorithm on these features. Third is model-based approaches where time series are modelled before being clustered. Another important choice is the distance measure like Euclidean, Kullback-Leibler divergence, Dynamic Time Warping (see Berndt and Clifford (1994)), etc.

Wavelets Decomposition (see Beylkin et al. (1991)) have yet being used in time series studies because it allows to work on the different levels of frequencies of the signal and to denoise the signal. We apply the pre-processing of Cugliari et al. (2016) which successfully applies disaggregated load clustering to forecast load demand. After time series projection, authors compute relative contribution of each energy level. Assume that $(\phi, \psi_{j,k})$ is a Haar basis. A continuous signal can be approximated in a truncated Haar basis: $\hat{f}(t) = c_0\phi(t) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t)$, where $c_0, d_{j,k}$ are the decomposition coefficients obtained with Fast Wavelet Transform algorithm. Then, we define relative contribution of level j by

$rel_j = logit\left(\frac{||\mathbf{d_j}||_2^2}{\sum_{k=0}^{J-1}||\mathbf{d_k}||_2^2}\right)$, where $logit(p) = \ln\left(\frac{p}{1-p}\right)$. In these features, we do not consider c_0, which corresponds to the mean level of each load. We focus our effort on the profiles form which is important to establish new pricing.

As in Cugliari et al. (2016), we use the relative contribution after a Haar Decomposition. In place of K-Means we use Maximum Distance to Average Vector generic (MDAV-generic) presented by Domingo-Ferrer and Torra (2005), because we want to have clusters of minimum size k regardless the number of clusters, and not k clusters regardless their size. Instead of MDAV-generic, we could have use less rigid algorithms like V-MDAV (see Solanas and Ballesté (2006)), which does not force each cluster (except some last) to be of a fixed size. Through MDAV-generic, we know in advance the number of clusters, which can be interested when there are a data furniture requirements specitication with the third party. We benchmark the technique with a mean based aggregation and a variance based aggregation. By mean (resp. variance) based aggregation, we mean achieving K Nearest Neighbor on the mean (resp. variance) load of each individual with initializing by the smallest mean (resp. variance). The benchmarks have some advantages : they are easy to implement and timely computated.

Cluster algorithm allows to divide the individuals in subsets of pre-determined minimum length. Then, we have to choose the aggregation techniques. We can compute the median load of the individuals of each cluster at each time. This choice allows to minimize the absolute loss for each cluster and to hide some extremes values. However, there is a non-zero probability that one individual of one cluster is (almost) always the median. In this case, giving the median is equivalent to give the load of one individual. The mean can be computed at each instant for each cluster. However, the attacker has access for each individual i to $\mathbf{A_i}$ the vector of daily load, and $(l_j)_j$ the infra-daily loads of each final aggregat (and so to (L_j) the daily loads of each cluster). The attacker can try to solve for each cluster j,

$$\arg\min_{p_i\in\{0,1\}} ||\mathbf{L_j} - \frac{1}{\sum p_i}\sum p_i\mathbf{A_i}||_2.$$

This adversaries model is then equivalent to a Knapsack problem (see Kellerer et al. (2004)), which can be solved by many algorithms (e.g. programming dynamic or simulated annealing). Even it is a NP-hard problem and the attacker needs an exact solution, many works show it is possible to consider the problem in a multi-parallel way and use GPU programming (see Boyer et al. (2012), Suri et al. (2012)). Adding a noise, even small, allows to get out of the knapsack problem.

Algorithm 1 MDAV-generic

Assume D the relative contribution dataset and k an integer.

1. While $card(D) \geq 3k$
   (a) Compute the average attribute-wise of all records in D
   (b) Compute the most distant record d_1 of previous average in term of Euclidian norm
   (c) Find the most distant record d_2 of d_1
   (d) Use d_2 and d_1 as the center of two clusters of length k
   (e) Delete the records of the two clusters and come back at the beginning
2. If $2k \leq card(D) \leq 3k-1$
   (a) Compute the average attribute-wise of remaining records in D
   (b) Compute the most distant record d_1 of previous average (Euclidian norm)
   (c) Use d_1 as the center of a cluster of length k
   (d) Form another cluster with the others remaining records in D
3. If $card(D) < 2k$, form a cluster with remaining records

5.3 Results

We compare the performance when clustering algorithm is applied on 1 400 centralized simulated time series. Remind the objectives consist in furniture of profil as homogeneous as possible to a third party which wants to propose new pricing. That means the third party has to collect cluster homogeneous, and data utility measure has to concern this point. If data are giving for a task of forecasting, we should measure differently the utility, for instance by computing MAPE (Mean

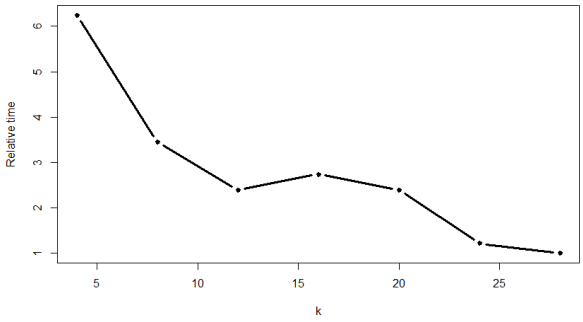

Figure 2: Process relative time

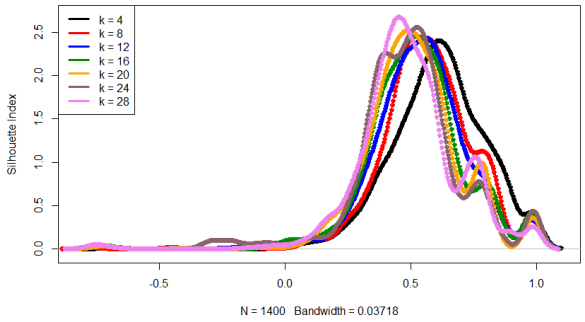

Figure 3: Silhouette Index estimated density

Absolute Percentage Error) on a test subset (e.g. see Pierrot and Goude (2011)). Here, in a task of pricing, there is no forecasting need. It illustrates the dependency between anonymization and future use of data. We measure data utility by computing indicators like silhouette index and the Davies-Bouldin index. These two indicators represent measures of homogeneity of clusters. We do not use indicators like Root Mean Square Error, because there are different load levels. We work from 4 to 28 anonymization by step of 4. In Figure 2, we plot the relative computation time when the reference level is the computation time of 28-anonymization. Computation time decreases when k grows. The Figure 3 gives an estimation of Silhouette Index density for each k-anonymization. This index, computed for each individual, is between -1 and 1. Stronger is this index, stronger the individual is connected to its cluster and far away the others clusters. It has to be upper than 0 if an individual is well clustered. Obviously, when k grows (data protection increases), the index decreases (data utility decreases). We see three modes in the density : the upper corresponds to second home, the medium to professional and the last to household.

Bouldin-Davies index represents the average similarity between each cluster and its most similar one, averaged over all the clusters. Lower this index is, the better the clustering is. In Figure 4, we give the ratio between Bouldin-Davies Index of each k-anonymization and the one of 28-anonymization. Figure 4 illustrates data utility loss caused by increasing data protection. There is a big gap of data utility between 4-anonymization and 8-anonymisation. However, the data protection is insufficient.

In Figure 5 we compare the MDAV-generic ap-

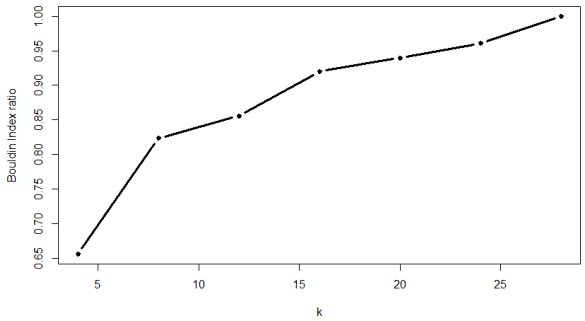

Figure 4: Bouldin-Davies Index ratio

plied on relative contribution of wavelet decomposition with two aggregations : one done by mean level (before centralization) and the other by standard deviation level. In relative Bouldin-Davies we give the ratio of Bouldin-Davies Index for each k-anomyzation by mean and standard deviation and the Bouldin-Davies Index when MDAV-generic is applied with the same k. All the indicators show MDAV-generic applied on relative contribution of wavelet decomposition outperforms the two forms of trivial aggregation, and second order (based on variance) aggregation is more efficient than first order (based on mean).

6 Conclusion

The process allows to formalize a global data-driven anonymization process facilizing the separation between specific and generic part of anonymization and integrating business knowledge and Data Science algorithms. Through this process formalization, we optimize the trade-off privacy protection/data utility.

To illustrate the process, we simulate a context near (but different) the situation of electrical

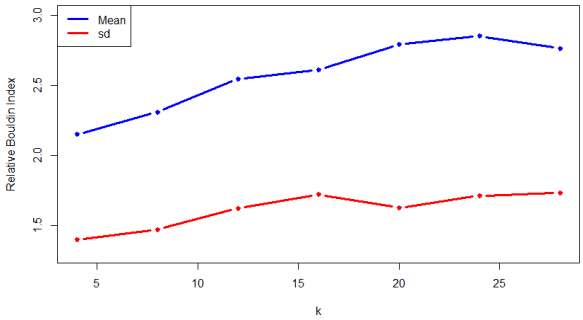

Figure 5: Relative Bouldin-Davies Index between benchmarks and MDAV-generic

smart meters data provision. We assume a third party tries making new pricing and propose a microaggregation process of time series using preprocessing through the methodology of Cugliari et al. (2016) and clustering algorithm of Domingo-Ferrer and Torra (2005). Instead punctual information giving at each instant, it could be interesting to give a probabilistic view load.

Our example is based on static data. In many applications, it is interesting to receive data in almost streaming way. One of the next step is to develop incremental microaggregation and measure the privacy loss and the data utility in this context. Another future work is the development of a big data framework allowing anonymization with noise addition, differential privacy and scalable microaggregation dividing the specific part inherent at each type of data, business constraint and future data utility with generic part. Lastly, here, we work of global datalake anonymization, which assumes there is a level where raw data are stored before transformation. Local anonymization, where data are anonymized at individual level have to be studied.

References

C.C. Aggarwal and P.S. Yu. 2008. A survey of randomization methods for privacy-preserving data mining. In *Privacy-Preserving Data Mining - Models and Algorithms*, pages 137–156.

G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. 2006. Achieving anonymity via clustering. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, ACM, New York, NY, USA, PODS '06, pages 153 – 162. https://doi.org/10.1145/1142351.1142374.

R. Agrawal and R. Srikant. 2000. Privacy preserving data mining. In *ACM SIGMOD Conference*.

M. Bergeat, N. Cuppens-Boulahia, F. Cuppens, N. Jess, F. Dupont, S. Oulmakhzoune, and G. De Peretti. 2014. A French Anonymization Experiment with Health Data. In *PSD 2014 : Privacy in Statistical Databases*. Eivissa, Spain. https://hal.archives-ouvertes.fr/hal-01214624.

D.J. Berndt and J. Clifford. 1994. Using dynamic time warping to find patterns in time series. In Usama M. Fayyad and Ramasamy Uthurusamy, editors, *KDD Workshop*. AAAI Press, pages 359–370.

G. Beylkin, R. Coifman, and V. Rokhlin. 1991. Fast wavelet transforms and numerical algorithms I. *Comm. Pure Appl. Math.* 44(2):141–183. https://doi.org/10.1002/cpa.3160440202.

K. Blazakis, S. Davarzani, G. Stavrakakis, and I. Pisica. 2016. Lessons learnt from mining meter data of residential consumers. *Periodica Polytechnica Electrical Engineering and Computer Science* 60(4):266–272. https://doi.org/10.3311/PPee.9993.

V. Boyer, D. El Baz, and M. Elkihel. 2012. Solving knapsack problems on {GPU}. *Computers and Operations Research* 39(1):42 – 47. Special Issue on Knapsack Problems and Applications. https://doi.org/http://dx.doi.org/10.1016/j.cor.2011.03.014.

E. Buchmann, K. Böhm, T. Burghardt, and S. Kessler. 2013. Re-identification of smart meter data. *Personal and Ubiquitous Computing* 17(4):653–662. https://doi.org/10.1007/s00779-012-0513-6.

J.W. Byun, A. Kamra, E. Bertino, and N. Li. 2007. *Efficient k-Anonymization Using Clustering Techniques*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 188–200.

K. Chatzikokolakis, C. Palamidessi, and M. Stronati. 2013. A predictive differentially-private mechanism for mobility traces. *CoRR* abs/1311.4008. http://arxiv.org/abs/1311.4008.

S. Chester, B. M. Kapron, G. Srivastava, and S. Venkatesh. 2013. Complexity of social network anonymization. *Social Netw. Analys. Mining* 3(2):151–166.

J-X. Chin, T. Tinoco De Rubira, and G. Hug. 2016. Privacy-protecting energy management unit through model-distribution predictive control. *CoRR* abs/1612.05120. http://arxiv.org/abs/1612.05120.

J. Cugliari, Y. Goude, and J. M. Poggi. 2016. Disaggregated electricity forecasting using wavelet-based clustering of individual consumers. In *2016 IEEE International Energy Conference (ENERGYCON)*. pages 1–6. https://doi.org/10.1109/ENERGYCON.2016.7514087.

G. Danezis. 2013. Privacy technology options for protecting and processing utility reading.

J. Domingo-Ferrer and V. Torra. 2005. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery* 11(2):195–212. https://doi.org/10.1007/s10618-005-0007-5.

F. Dufaux and T. Ebrahimi. 2006. Scrambling for video surveillance with privacy. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. pages 160–160. https://doi.org/10.1109/CVPRW.2006.184.

C. Dwork. 2006. *Differential Privacy*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 1–12.

C. Dwork and A. Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9(3Ű4):211–407. https://doi.org/10.1561/0400000042.

K. Ezhilarasi and Mr. M. Hariharan. 2015. Protecting sensitive information in bank transaction with data anonymization. *International Journal For Technological Research In Engineering* (2 (11)).

S. Fan and R.J. Hyndman. 2012. Short-term load forecasting based on a semi-parametric additive model. *Power Systems, IEEE Transactions on* 27(1):134–141. https://doi.org/10.1109/TPWRS.2011.2162082.

P. Gaillard, Y. Goude, and R. Nedellec. 2016. Additive models and robust aggregation for gefcom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting* 32(3):1038–1050.

S. Gambs, M-O. Killijian, and M. Nunez del Prado Cortez. 2014. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences* 80(8):1597 – 1614. Special Issue on Theory and Applications in Parallel and Distributed Computing Systems.

B. Gedik and L. Liu. 2008. Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *IEEE Transactions on Mobile Computing* 7(1):1–18. https://doi.org/10.1109/TMC.2007.1062.

S. Hansell. 2006. Aol removes search data on vast group of web users. *New York Times* .

T. J. Hastie and R. J. Tibshirani. 1990. *Generalized additive models*. London: Chapman & Hall.

S-K. Hong, K. Gurjar, H-S. Kim, and Y.S. Moon. 2013. A survey on privacy preserving time-series data mining. *3rd International Conference on Intelligent Computational Systems (ICICS'2013)* .

T. Hong, P. Pinson, and S. Fan. 2014. Global energy forecasting competition 2012. *International Journal of Forecasting* 30(2):357 – 363.

H. Kellerer, U. Pferschy, and D. Pisinger. 2004. *Introduction to NP-Completeness of knapsack problems*. Springer.

W. Labeeuw and G. Deconinck. 2013. Residential electrical load model based on mixture model clustering and markov models. *IEEE Transactions on Industrial Informatics* 9(3):1561–1569. https://doi.org/10.1109/TII.2013.2240309.

N. Li, T. Li, and S. Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. pages 106–115. https://doi.org/10.1109/ICDE.2007.367856.

Y. Li, S. Zhu, L. Wang, and S. Jajodia. 2002. *A Privacy-Enhanced Microaggregation Method*, Springer Berlin Heidelberg, Berlin, Heidelberg, pages 148–159.

T. Warren Liao. 2005. Clustering of time series data, a survey. *Pattern Recognition* 38:1857–1874.

J-L Lin and M-C Wei. 2008. An efficient clustering method for k-anonymization. In *Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, PAIS 2008*. pages 499–502. https://doi.org/10.1109/CANDAR.2015.61.

K. Liu, C. Giannella, and K. Kargupta. 2008. A survey of attack techniques on privacy-preserving data perturbation methods.

G. Loukides and J-H. Shao. 2008. An efficient clustering algorithm for k-anonymisation. *Journal of Computer Science and Technology* 23(2):188–202. https://doi.org/10.1007/s11390-008-9121-3.

C. Y.T. Ma and D. K.Y. Yau. 2015. On information-theoretic measures for quantifying privacy protection of time-series data. In *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*. ACM, New York, NY, USA, ASIA CCS '15, pages 427–438. https://doi.org/10.1145/2714576.2714577.

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. 2006. l-diversity: Privacy beyond k-anonymity. In *IN ICDE*.

F. McLoughlin, A. Duffy, and M. Conlon. 2010. The generation of domestic electricity load profiles through markov chain modelling. *Euro-Asian Journal of Sustainable Energy Development Policy* 3.

D.HO. McQueen, P. R. Hyland, and S. J. Watson. 2004. Monte carlo simulation of residential electricity demand for forecasting maximum demand on distribution networks. *IEEE Transactions on Power Systems* 19(3):1685–1689.

F. McSherry and I. Mironov. 2009. Differentially private recommender systems: building privacy into the net. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, KDD '09, pages 627–636. https://doi.org/10.1145/1557019.1557090.

M. Muratori, Ma. C. Roberts, R. Sioshansi, V. Marano, and G. Rizzoni. 2013. A highly resolved modeling technique to simulate residential power demand. *Applied Energy* 107:465 – 473.

A. Narayanan and V. Shmatikov. 2008. Robust deanonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. pages 111–125. https://doi.org/10.1109/SP.2008.33.

R. Nedellec, J. Cugliari, and Y. Goude. 2014. Gefcom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting* 30(2):375 – 381.

J. Nin and V. Torra. 2009. Towards the evaluation of time series protection methods. *Information Sciences* 179(11):1663 – 1677. Including Special Issue on Chance Discovery.

J.V. Paatero and P.D. Lund. 2005. A model for generating household electricity load profiles. *International journal of energy research* 30.

A. Pierrot and Y. Goude. 2011. Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP power* pages 593–600.

P. Pompey, A. Bondu, Y. Goude, and M. Sinn. 2015. *Massive-Scale Simulation of Electrical Load in Smart Grids Using Generalized Additive Models*, Springer International Publishing, Cham, pages 193–212.

S. Rani and G. Sikka. 2012. Recent techniques of clustering of time series data: A survey. *International Journal of Computer Applications* .

E. Shi, T.-H. H. Chan, E. G. Rieffel, R. Chow, and D.Song. 2011. Privacy-preserving aggregation of time-series data. In *Proceedings of the Network and Distributed System Security Symposium, NDSS 2011, San Diego, California, USA, 6th February - 9th February 2011*.

L. Shou, X. Shang, K. Chen, G. Chen, and C. Zhang. 2013. Supporting pattern-preserving anonymization for time-series data. *IEEE Transactions on Knowledge and Data Engineering* 25(4):877–892. https://doi.org/10.1109/TKDE.2011.249.

A. Solanas and A. Ballesté. 2006. V-mdav: Variable group size multivariate microaggregation .

B. Suri, U. D. Bordoloi, and P. Eles. 2012. A scalable gpu-based approach to accelerate the multiple-choice knapsack problem. In *2012 Design, Automation Test in Europe Conference Exhibition (DATE)*. pages 1126–1129. https://doi.org/10.1109/DATE.2012.6176665.

L. Sweeney. 2002. K-anonimity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(05):557–570. https://doi.org/10.1142/S0218488502001648.

V. Torra. 2017. *Data Privacy: Foundations, New Developments and the Big Data Challenge (Studies in Big Data)*. Springer.

G. Tóth, Z. Hornák, and et al. 2004. Measuring anonymity revisited.

V. Tudor, M. Almgren, and M. Papatriantafilou. 2013. Analysis of the impact of data granularity on privacy for the smart grid. *WPES '13 Proceedings of the 12th ACM* .

S. Venkatasubramanian. 2008. *Measures of Anonymity*, Springer US, Boston, MA, pages 81–103.

X. Xu and M. Numao. 2015. An efficient generalized clustering method for achieving k-anonymization. In *2015 Third International Symposium on Computing and Networking (CANDAR)*. pages 499–502. https://doi.org/10.1109/CANDAR.2015.61.

G. Zhang, X. Liu, and Y. Yang. 2015. Time-series pattern based effective noise generation for privacy protection on cloud. *IEEE Transactions on Computers* 64(5):1456–1469. https://doi.org/10.1109/TC.2014.2298013.

Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. 2007. Aggregate query answering on anonymized tables. In *2007 IEEE 23rd International Conference on Data Engineering*. pages 116–125. https://doi.org/10.1109/ICDE.2007.367857.

B. Zhou, J. Pei, and W. Luk. 2008. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explor. Newsl.* 10:12–22. https://doi.org/10.1145/1540276.1540279.