DOCUMENT DE TREBALL

XREAP2018-2

DETECTING OUTLIERS WITH SEMI-SUPERVISED MACHINE LEARNING: A FRAUD PREDICTION APPLICATION

Sebastián M. Palacio (GiM, XREAP)

XARXA DE REFERÈNCIA EN ECONOMIA APLICADA

Electronic copy available at: https://ssrn.com/abstract=3165318

Detecting Outliers with Semi-Supervised Machine Learning: A Fraud Prediction Application

Sebastián M. Palacio* spalacpa9.alumnes@ub.edu
Department of Econometrics, Statistics and Applied Economics
Barcelona University
Barcelona, Diagonal 690, Spain

Editor:

Abstract

Abnormal pattern prediction has received a great deal of attention from both academia and industry, with applications that range from fraud, terrorism and intrusion detection to sensor events, medical diagnoses, weather patterns, etc. In practice, most abnormal pattern prediction problems are characterized by the presence of a small number of labeled data and a huge number of unlabeled data. While this points most obviously to the adoption of a semi-supervised approach, most empirical studies have opted for a simplification and treated it as a supervised problem, resulting in a severe bias of false negatives. In this paper, we propose an innovative methodology based on semi-supervised techniques and introduce a new metric the Cluster-Score for abnormal homogeneity measurement. Specifically, the methodology involves transmuting unsupervised models to supervised models using the Cluster-Score metric, which defines the objective boundaries between clusters and evaluates the homogeneity of the abnormalities in the cluster construction. We apply this methodology to a problem of fraud detection among property insurance claims. The objectives are to increase the number of fraudulent claims detected and to reduce the proportion of claims investigated that are, in fact, non-fraudulent. The results from applying our methodology considerably improved these objectives.

Keywords: Outlier Detection, Semi-Supervised Models, Fraud, Cluster, Insurance

1. Introduction

The problem we seek to solve is the prediction of abnormalities in an environment with highly unbalanced samples and a huge mass of unlabeled data. A typical example of such a situation is provided by fraud detection. In general, we only have partial information about fraud cases, as well as possibly some information about false positives, that is, cases that are considered suspicious but which prove to be cases of non-fraud. The problem here is that we cannot label these cases non-fraud simply because they were initially considered suspicious. For this reason, we know nothing about non-fraud cases. Moreover, fraud tends to be an outlier problem, given that we are dealing with atypical values with respect to regular data. Hence, it is likely that we only dispose of information about an extremely small sample. Yet, it so transpires, that this information is extremely useful and should not be discarded. In contrast we have a considerable amount of data that may contain fraud

∗. The author would like to thank to Cristina Rata and Joan-Ramon Borrell for constructive criticism of the manuscript

Electronic copy available at: https://ssrn.com/abstract=3165318

and or non-fraud cases and, as such, we cannot treat these data using traditional supervised algorithms.

The problem, simply stated, therefore, is how can we predict these outliers? To represent this typical case we apply an innovative semi-supervised methodology to a real fraud case. Specifically, we draw on information provided by a leading insurance company as we seek to predict fraudulent insurance claims. In general terms, such claims fall into two categories: one, those that provide only partial or untruthful information in the policy contract; and, two, those that are based on misleading or untruthful circumstances (including exaggerations). It has been estimated that cases of detected and undetected fraud represent up to 10% of all claims in Europe (The Impact of Insurance Fraud, 2013), accounting for around 10-19% of the payout bill.

In the sector, the main services contracted are automobile and property insurance, representing 76% of total claim costs. However, while many studies have examined automobile fraud detection (see, for example, Arts et al., 1999 and 2002; Viaene et al., 2007; Wilson, 2009; Nian et al., 2016), property fraud has been largely neglected, perhaps because detection is more difficult as witnesses are infrequent or typically tend to cohabitants.

Here, therefore, our main objective is to present a variety of semi-supervised machine learning models applied to a fraud detection problem. In so doing, we aim to develop a methodology capable of improving results in classification anomaly problems of this type. Our reasoning for using semi-supervised models is best explained as follows. Statistically speaking, fraud is a special case of outliers, that is, of points in the dataset that differ significantly from the remaining data. Such anomalies often result from unusual events that generate anomalous patterns of activity. Were we to use unsupervised models that is, were we to assume that we are unable to distinguish between fraudulent and non-fraudulent cases what we defined as outliers, noise or normal data would be subjective and we would have to represent that noise as a boundary between normal data and true anomalies without any information. But, as mentioned, the number of fraud cases detected is small; however, they constitute a useful source of information that cannot be discarded.

On the other hand, supervised models are inappropriate because, in general, we face a major problem of claim misclassifications when dealing with fraud detection (Arts et al., 2002). Fraud detection, typically, comprises two stages: first, it has to be determined whether the claim is suspicious or not (Viaene et al, 2007); and, second, all cases considered suspicious have to be examined by fraud investigators to determine whether the claim is fraudulent or not. This means that unsuspicious cases are never examined, which is reasonable in terms of efficiency, especially if the process cannot be automatized. Insurance adjusters have little time to perform an exhaustive investigation. Yet, the process does provide us with partial information, that is, labels for what is a small sample. Clearly, using a supervised model in this instance adds bias to the confusion matrix. Essentially, we will detect severe bias in false negatives and, therefore, many cases which are in fact fraudulent will be predicted as being non-fraudulent (Phua et al., 2004). Indeed, when using supervised algorithms we assume that the system in place is capable of discerning perfectly between fraudulent and non-fraudulent claims, an outcome that in practice is infrequent and referred to in the literature as an omission error (Bollinger and David, 1997; Poterba and Summers, 1995).

Clearly, the information provided in relation to those cases considered suspicious is more likely to be specified correctly once we have passed the first stage in the fraud detection

process. This information will be useful for a part of the distribution (that is, it will reveal if a fraudulent claim has been submitted), which is why it is very important this information be taken into account. For this reason, fraud detection is notorious for being considered a semi-supervised problem because the ground truth labelling of the data is partially known. Here, therefore, we seek to make three contributions to the literature: First, we apply innovative semi-supervised techniques to anomaly detection; second, we create a new metric that permits us to evaluate the homogeneity of abnormalities in the cluster construction; and, third, we apply this model to an actual property insurance claim fraud problem for the first time, using a real dataset provided by a leading insurance company.

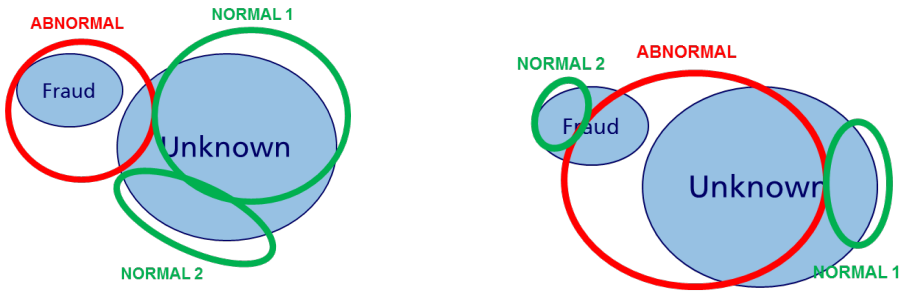
2. Metodhology

Outlier detection models seek to separate regular from outlier observations. If we have labeled data, the easiest way to proceed is by employing a supervised algorithm. However, in the case of fraud, this implies our knowing everything about the two classes of observation that is, we would know exactly who did and did not commit fraud, a situation that is extremely rare. In contrast, if we know nothing about the labeling, that is, we do not know who did and did not commit fraud, several unsupervised methods of outlier detection can be employed, for example, isolation forest, one-class support vector machines, and elliptic envelopment. However, they tend to be less precise and we have to assume some subjective boundary.

If, however, we have some label data about each class, we can implement a semi-supervised algorithm, such as label propagation or label spreading. Yet, these methods require that we have some information about every class in our problem, something that is not always possible. Indeed, disposing of label data information about each class is quite infrequent in certain practical problems. Additionally, we face the problem of unbalanced data, which means we rarely have clean, regular data representing the population. In fraud problems, as a norm, the data are highly imbalanced and skewed, which results in a high but biased success rate.

In the light of these issues, we propose an innovative semi-supervised technique that can assess not only a highly unbalanced dataset problem but also one for which we have no information about certain classes. In this regard, fraud detection represents an outlier problem for which we can usually identify some, but not all, of the cases. We might, for example, have information about false positives, that is, investigated cases that proved not to be fraudulent. However, simply because they have raised suspicions does not mean they can be considered representative of non-fraudulent cases. In short, what we usually have are some cases of fraud and a large volume of unknown cases (among which it is highly likely cases of fraud are lurking).

Bearing this in mind, we propose the application of unsupervised models so as to relabel the target variable. To do this, we use an innovative metric that measures how well we approximate the minority class. We can then transform the model to a semi-supervised algorithm. On completion of the relabeling process, our problem can be simplified to a supervised model. This allows us not only to set an objective boundary but to obtain a



(a) Desired Clusters (b) Non-desired Clusters

Fig. 1. Possible clusters

gain in accuracy when using partial information, as Trivedi et al. (2014) have demonstrated.

2.1 Unsupervised Model Selection

We start with a dataset of 303,166 cases. We set aside a 10% random subset for final evaluation. Hence, our dataset consists of 270,479 non-identified cases and 2,370 cases of fraud. The main problem we face in this unsupervised model is having to define a subjective boundary. We have partial information about fraud cases, but have to determine an acceptable threshold at which an unknown case can be considered fraudulent. When calculating unsupervised classification models, we reduce the dimensions to clusters. Almost every algorithm will return several clusters containing mixed-type data (fraud and unknown). Intuitively, we would want the fraud points revealed to be highly concentrated into just a few clusters. Likewise, we would expect some non-revealed cases to be included with them, as in Figure 1.a:

On the other hand, we would want to avoid situations in which abnormal and normal cases are uniformly distributed between groups, as in Figure 1.b. However, a limit of some kind has to be defined. But, how many of the unknown cases can we accept as being fraudulent?

A boundary line might easily be drawn so that we accept only cases of detected fraud or we accept every possible case as fraudulent. Yet, we know this to be unrealistic. If we seek to operate between these two both extremes, intuition tells us that we need to stay closer to the lower threshold, accepting only cases of fraud and very few more, as Figure 2 illustrates.

But once more, we do not know exactly what the correct limit is. In this way, however, we have created an experimental metric that can help us assign a score and, subsequently, define the threshold. This metric, which we shall refer to as the cluster score (CS), calculates the weighted homogeneity of clusters based on the minority and majority classes.

$$CS \;=\; (1+\alpha^2)\frac{C1 * C2}{C1 + C2 * \alpha}$$

Essentially, it assigns a score to both the minority-class (C1) and the majority-class (C2) clusters based on the weighted conditional probability of each point. The mathematical

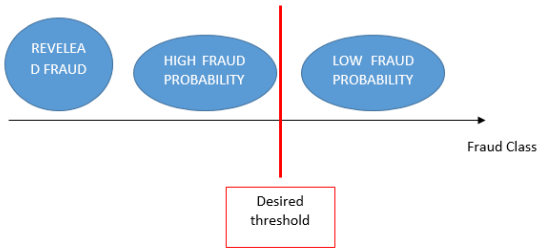
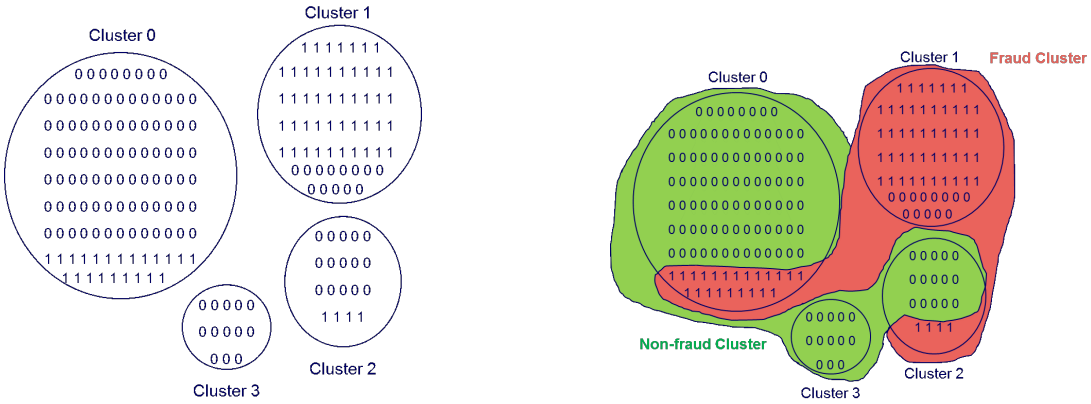


Fig. 2. Desired threshold



(a) Unsupervised output

(b) Relabeled clusters

Fig. 3. Possible clusters

formulation can be consulted in Appendix 1.

We should stress that each time we retrieve more information about the one-class cases that have been revealed, this threshold improves. This is precisely where the entropy process of machine learning appears. In the one-class fraud problem discussed above, we start with an unknown distribution for which some data points are known (that is, the fraud sample). Our algorithms, using the CS proposed, will gradually get closer to the best model that can fit these cases of fraud, while maintaining a margin for undiscovered cases. Now, if we obtain new information about fraud cases, our algorithms will readjust to provide the max CS again. As the algorithms work with notions based on density and distances, they change their shapes to regularize this new information.

Once the best unsupervised model is attained (that is, the model that reaches the max CS), we need to decide what to do with the clusters generated. Basically, we need to determine which clusters comprise fraudulent and which comprise non-fraudulent cases. The difficulty is that several cluster will be of mixed-type: that is, minority-class points (fraud cases) and unidentified cases, as in Figure 3.a, where the 0s are unidentified cases and the 1s are minority-class points.

In defining a threshold for a fraud case, we make our strongest assumption. Here, we assume that if a cluster is made up of more than 50% of fraud cases, this cluster is a

fraud cluster, otherwise, it is a non-fraud cluster. The distinction introduced is clear: The non-fraud cluster is no longer an unidentified cluster. By introducing this assumption, we state that they are actually non-fraudulent cases. This definition acts as the key for our transition into a semi-supervised model.
As Figure 3.b shows, cluster 1, being composed of more than 50% fraud cases, now forms part of the more general fraud cluster, together, obviously, with the fraud cases already detected. The remaining cases that do not belong to such a dense fraud cluster are now considered non-fraud cases.
The unsupervised algorithm can thus be summarized as follows:

Algorithm 1 Unsupervised Algorithm

1. Load dataset: Overweight fraud cases given the same weight as unknown cases.
2. Iterate through model k and the vector of tuning parameters$i$.
3. j clusters will generate for each $k, i.$ that is $C_{i,k} = \{C^1_{i,k}, C^2_{i,k} C^n_{i,k}\}$.
4. For each $C^j_{i,k}$ we calculate $C1$ and $C2$ (see Appendix), and obtain the cluster score $CS_{i,k}$.
5. We define the acceptable threshold t^* for a cluster to be considered a fraud cluster or not.
6. We choose the optimal CS^* where $CS^* = argmax\{CS_{I,K}\}$, $CS_{I,K}$ is the cluster vector for each pair k, i.
7. We relabel the fraud variable using the optimal clustering model derived from CS^*. Each unknown case in a fraud cluster is now equal to 1. Known fraud cases are equal to 1. Remaining cases are equal to 0.

2.2 Supervised Model Selection

We now have a redefined target variable that we can continue working with by applying an easy-to-handle supervised model. The first step involves resampling the fraud class to avoid unbalanced sample problems. We oversample the dataset to obtain a 50/50 balanced sample.
The second step involves conducting a grid search and a ten cross-validation (CV) based on the F-Score[1] to obtain the optimal parameters for three different models: extreme randomized tree, gradient boosting and a light XGB. Additionally, we combine these classifiers using stacking models.
Once we have the optimal parameters for each, we calculate the optimal threshold that defines the probability of a case being fraudulent or non-fraudulent, respectively.

1. The F-Score was constructed using $\beta = 2$, as we needed to place greater weight on the recall.

Finally, we identify the two models that perform best on the combined valid/test dataset the best acting as our main model implementation, the other controlling that the predicted claims are generally consistent.

Algorithm 2 Supervised Algorithm

1. Load dataset with the relabeled target variable. Oversampling to obtain a 50/50 balance sample.
2. Apply grid search and CV to obtain the optimal parameters based on the F-Score results for each M_i model.
3. Generate stacking models S_j using different combinations of metaclassifiers based on the M_i models.
4. Obtain the optimal threshold for each M_i and S_j based on F-Score results.
5. Identify the best two models. The best model is Base Model B^*. And the second best is Control Model C^*.
6. Finalize the output probabilities.

3. Data

We use an insurance fraud dataset provided by a leading insurance company in Spain for the period 2015-2016. After sanitization, our main sample consists of 303,166 property claims, some of which have been analyzed as possible cases of fraud by the Investigation Office (IO)[2].
Of the cases analyzed by the IO, 48% proved to be fraudulent. A total of 2,641 cases were resolved as true positives (0.8% of total claims) during the period under study. This means we do not know which class the remaining 99.2% of cases belong to. However, the fraud cases detected provide very powerful information, as they reveal the way in which fraudulent claims behave. Essentially, they serve as the pivotal cluster for separating normal from abnormal data.
A data lake was constructed during the process to generate sanitized data. We obtain 19 bottles containing different types of information related to accident claims. These bottles contain variables derived from the companys daily operation, and variables that are transformed in several of their aspects. In total we have almost 1,000 variables. We briefly present them here to help explain which concepts were chosen for the model.

4. Results

Table 2 shows the main unsupervised modeling results. Mixed models tend to provide the best results. However, mini-batch K-means is not only much faster, it also provides the

2. The system applied before to detect fraud corresponds to a rule based methodology.

Bottles	Descriptions
Id	ID about accident, policy, person, etc.
Customer	Policyholders attributes embodied in insurance policies: name, sex, age, address, etc.
Customer-Property	Customer related with the property data
Dates	Dates of about accident, policy, visits, etc.
Guarantees	Coverage and guarantees of the subscribed policy
Property	Data related to the insured object
Payments	Policy payments made by the insured
Policy	Policy contract data, including changes, duration, etc.
Loss-Adjuster	Information about the process of the investigation but also about the loss adjuster
Accident	Brief, partial information about the accident, including date and location
Intermediary	Information about the policies intermediaries.
Customer-Object-Reserve	The coverage and guarantees involved in the accident
Historical Accident	Historical movements associated with the reference accident
Historical Policy	Historical movements associated with the reference policy (the policy involved in the accident).
Historical Other Policies	Historical movements of any other policy (property or otherwise) related to the reference policy.
Historical Other Accident	Historical accident associated with the reference policy (excluding the accident analyzed).
Historical Other Policy-Accident	Other accident associated with other policies not in the reference policy (but related to the customer).
Black List	Every participant involved in a fraudulent claim (insured, loss-adjuster, intermediary, other professionals, etc).
Cross Variables	Several variables constructed with the interaction between the bottles.

Table 1. Data Bottles

best results, returning four clusters. More than 90% of the cases in the central cluster are fraudulent (well above our 50% threshold), but it also contains an additional 8,540 unknown cases. This is our core fraud cluster and the one we use when renaming the original labels. We also have a small fraud cluster with an additional 16 cases. C1 indicates that the minority-class (fraud) clusters comprise 92.9% of minority data points on a weighted average. In contrast, C2 indicates they are made up of 92.8% of unknown cases.

Model	Parameter 1	Parameter 2	n Clusters	C1	C2	Cluster Score	Time
Mini-Batch K-Means	Iterations= 100	Batch Size = 600	4	92.9%	92.8%	92.8%	3 sec.
Isolation Forest	Contamination= 0.125	$n_estim = 100$	2	51.5%	51.1%	51.4%	5min-18min
DBSCAN	Does not change	Does not change	2	50.2%	49.8%	50.1%	15-80min
Gaussian Mixture	Cov= *Tied*	Tol= 0.29	5	95.0%	95.0%	96.3%	23min
Bayesian Mixture	Cov= *Tied*	Tol= 0.29	6	96.5%	96.4%	96.5%	23min

Table 2. Unsupervised model results

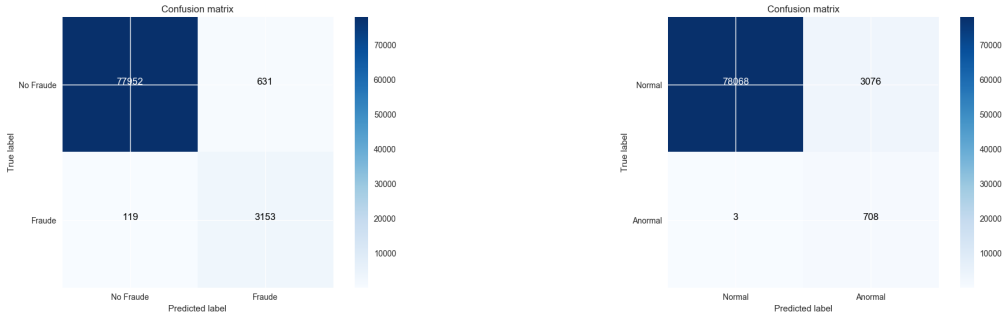
The tuning of each of the supervised models and of the stacking models is shown in Table 3. We first employ a validation set comprising 15% of the data to tune the parameters. As can be appreciated, we have two recall values. The cluster recall is the metric derived when using the relabeling target variable. The original recall emerges when we recover the prior labelling (1 if it was fraud, 0 otherwise). As can be seen, the results are strikingly consistent. We are able to predict fraud cluster with a recall of up to 91-98% in every case. But, more impressively yet, we can capture the original fraud cases with a recall close to 99%. The precision is a little lower, but in almost all cases it is higher than 80%. These are particularly good results for a problem that began as an unsupervised high-dimensional problem in an extremely unbalanced one-class dataset.

After optimizing the models, we calculated the performance of the best models using 70/30 training-test dataset. The performance and confusion matrixes are shown in Figure 4. Both are extremely randomized trees: the first uses balance subsampling (ERT-ss) and serves here as our base model; the second uses an ADASYN oversampling method (ERT

Model	Cluster Recall	Original Recall	Precision	FBeta	Treshold	Time
ERT b-ss	96.7%	99.4%	81.8%	93.3%	0.539	3min
ERT b-s	97.3%	99.1%	81.7%	93.7%	0.507	10min
GB	91.0%	94.6%	74.0%	87.0%	0.634	6h30min
LXGB	91.3%	94.6%	84.1%	89.8%	0.688	4min
Stacking META=ERT	97.3%	99.2%	49.9%	81.8%	0.955	12hs
Stacking META=LXGB	93.1%	95.2%	84.0%	91.2%	0.552	12hs
Stacking META=GB	91.9%	95.2%	84.7%	90.4%	0.661	15min

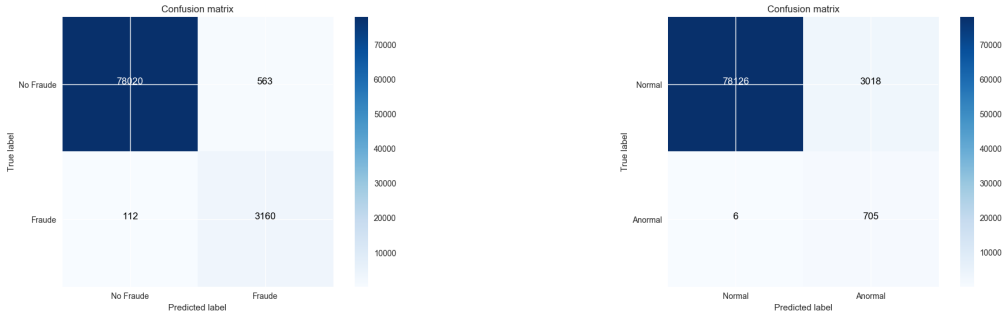
Table 3. Supervised model results

b-s) and serves as our control model. The results continue to be consistent and they are very rapid algorithms.



(a) Cluster target (b) Original target

Fig. 4. Base Model. ERT with subsampling



(a) Cluster target (b) Original target

Fig. 5. Control Model. ERT with oversampling

5. Robustness Check

At the outset, we randomly set aside 10% of the data (30,317 claims). In this final step, to ensure we do not face any overfitting problems, we examine these initial data in a test of the whole process. Our results are shown in Table 4.

Original Value	Prediction	Cases
NOT INVESTIGATED	NOT FRAUD	29.631
FRAUD	NOT FRAUD	0
NOT INVESTIGATED	FRAUD	415
FRAUD	FRAUD	271

(a) Base Model Robustness Check.

Original Value	Prediction	Cases
NOT INVESTIGATED	NOT FRAUD	29.656
FRAUD	NOT FRAUD	8
NOT INVESTIGATED	FRAUD	390
FRAUD	FRAUD	263

(b) Control Model Robustness Check.

Table 4. Model Robustness Check.

As can be appreciated, the control model (Table 4.b) has a recall of 97% while the base model (Table 4.a) has an impressive recall of 100%. However, the added value depends on the non-investigated fraud cases, that is, cases not previously detected but which would boost our results if shown to be fraudulent. We, therefore, sent these cases to the IO for analysis.

The IO investigated 367 cases (at the intersection between the control and base models). Two fraud investigators analyzed each of these cases, none of which they had previously seen as the rule model had not detected them.

Of these 367 cases, 333 were found to present a very high probability of being fraudulent. This means that only 34 could be ruled out as not being fraudulent. Recall that from the original sample of 415 cases, the fact that 333 presented indications of fraud means we have a precision of 88%, which is perfectly consistent with our original test sample (83.3-84.9%). In short, we managed to increase the efficiency of fraud detection by 122.8%. These final outcomes are summarized in Table 5.

Original Value	Prediction	Cases
NOT INVESTIGATED	NOT FRAUD	29.631
FRAUD	NOT FRAUD	0
NOT FRAUD	FRAUD	(415 - 333) = 82
FRAUD	FRAUD	(271 + 333) = 604

Table 5. Base Model Final Results

6. Machine Learning Process

The machine learning process is not complete if we do not feed the model with new information. A year later, we recalibrated the model with new data (see Table 6).

PERIOD	**Jan15-Jan17**	**Jan15-Jan18**
Claims	303,166	479,454
Observed Fraud	2,641	4,299
Cluster Score	92.8%	97.13%
Recall Score	96.4%	97.4%
Precision Score	83.3%	90.4%

Table 6. Base Model with the machine-learning process applied

The base model greatly improves the homogeneity of the fraud and non-fraud clusters. In particular, it provides a gain of 7% in the precision score and a slight gain in the recall score.

7. Conclusions

This paper has sought to offer a solution to the problems that arise when working with highly unbalanced datasets for which the labelling of all the classes is unknown. In such cases, we usually dispose of a few small samples that contain highly valuable information. Here, we have presented a fraud detection case, drawing on the data provided by a leading insurance company, and have tested a new methodology based on semi-supervised fundamentals to predict fraudulent accident claims.

At the outset, the IO did not investigate many cases (around 7,000 cases from a total of 303,166). Of these, only 2,641 were actually true positives (0.8% of total claims), with a success rate of 48%. Thanks to the methodology devised herein, we can now investigate the whole spectrum of cases automatically, obtaining a *total recall of 96-97%* and a precision of *83-90%*. In spite of the complexity of the initial problem, where the challenge was to detect outliers without knowing anything about 99.2% of the sample, the methodology described has been shown to be capable of solving the problem with great success.

References

Manuel Arts, Mercedes Ayuso, and Montserrat Guillen. Modelling different types of automobile insurance fraud behaviour in the spanish market. *Insurance: Mathematics and Economics*, 24(1-2):67–81, 1999. URL https://EconPapers.repec.org/RePEc:eee:insuma:v:24:y:1999:i:1-2:p:67-81.

Manuel Arts, Mercedes Ayuso, and Montserrat Guillen. Detection of automobile insurance fraud with discrete choice models and misclassified claims. 69:325 – 340, 09 2002.

Rekha Bhowmik. Detecting auto insurance fraud by data mining techniques. 2:156–162, 04 2011.

Christopher R. Bollinger and Martin H. David. Modeling discrete choice with response error: Food stamp participation. *Journal of the American Statistical Association*, 92 (439):827–835, 1997. URL http://www.jstor.org/stable/2965547.

Stephen Chiu. Fuzzy model identification based on cluster estimation. 2:267–278, 01 1994.

Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996. URL http://dl.acm.org/citation.cfm?id=3001460.3001507.

Kevin J. Leonard. Detecting credit card fraud using expert systems. 25:103–106, 09 1993.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY*, 49(4): 764–766, 2013. URL http://dx.doi.org/10.1016/j.jesp.2013.03.013.

Alexander Lindholm. A study about fraud detection and the implementation of suspect - supervised and unsupervised erlang classifier tool, 2014. ISSN 1401-5749.

F. T. Liu, K. M. Ting, and Z. H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 12 2008.

Ke Nian, Haofan Zhang, Aditya Tayal, Thomas Coleman, and Yuying Li. Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science*, 2(1):58–75, 2016. URL http://www.sciencedirect.com/science/article/pii/S2405918816300058.

Richard Oentaryo, Ee-Peng Lim, Michael Finegold, David Lo, Feida Zhu, Clifton Phua, Eng-Yeow Cheu, Ghim-Eng Yap, Kelvin Sim, Minh Nhut Nguyen, Kasun Perera, Bijay Neupane, Mustafa Faisal, Zeyar Aung, Wei Lee Woon, Wei Chen, Dhaval Patel, and Daniel Berrar. Detecting click fraud in online advertising: A data mining approach. *Journal of Machine Learning Research*, 15:99–140, 2014. URL http://jmlr.org/papers/v15/oentaryo14a.html.

Clifton Phua, Damminda Alahakoon, and Vincent Lee. Minority report in fraud detection: Classification of skewed data. *SIGKDD Explor. Newsl.*, 6(1):50–59, June 2004. doi: 10.1145/1007730.1007738. URL http://doi.acm.org/10.1145/1007730.1007738.

Yusuf Sahin and Ekrem Duman. Detecting credit card fraud by decision trees and support vector machines. 1:442–447, 03 2011.

Aastha Sharma, Setu K. Chaturvedi, Bhupesh Gour, Folorunsho Olaiya, Adesesan B. Adeyemo, Barnabas Adeyemo, Zahoor Jan, Muhammad Abrar, Shariq Bashir, and Yujie Zheng. A semi- supervised technique for weather condition prediction using dbscan and knn. 2014.

S. Viaene, R. A. Derrig, and G. Dedene. A case study of applying boosting naive bayes to claim fraud diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 16(5): 612–620, 05 2004.

Stijn Viaene, Mercedes Ayuso, Montserrat Guillen, Dirk Van Gheel, and Guido Dedene. Strategies for detecting fraudulent claims in the automobile insurance industry. *European Journal of Operational Research*, 176(1):565–583, 01 2007. URL https://ideas.repec.org/a/eee/ejores/v176y2007i1p565-583.html.

Masoumeh Zareapoor and Pourya Shamsolmoali. Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia Computer Science*, 48:679–685, 2015. URL http://www.sciencedirect.com/science/article/pii/S1877050915007103.

Appendix A. Cluster Score

C1 Score

The C1 score calculates the probability that a revealed fraud point belongs to group J and that this probability is weighted by the total number of fraud points in that group divided by the total cases of fraud revealed. Mathematically this can be expressed as:

$$C1 \;=\; \frac{\frac{\sum_{j=1\ldots n} x^j_{fraud}}{X^T_j} * x^j_{fraud}}{X^T_{fraud}} \in [0,1]$$

where x^j_{fraud} is the revealed fraud point in cluster group j. X^T_j is the total number of points in group J. And X^T_{fraud} is the total number of fraud cases revealed.

Basically, we calculate the probability of a point belonging to a specific group j ($\frac{\sum_{j=1\ldots n} x^j_{fraud}}{X^T_j}$) and we weight these values by the participation of its points in the fraud group revealed. Our objective is to maximize $C1$; basically, this means ensuring all cases of revealed fraud are in the same groups. The limit $C1 = 1$ implies that in the j groups we only have revealed fraud points. However, this is not what we want. Therefore, we have to balance this function with another function.

C2 Score

$C2$ is the inverse case of $C1$. Here, we calculate the probability of unrevealed fraud belonging to group J.

$$C2 \;=\; \frac{\frac{\sum_{j=1\ldots n} x^j_{not-fraud}}{X^T_j} * x^j_{not-fraud}}{X^T_{not-fraud}} \in [0,1]$$

And the objective is the same as that above in the case of $C1$: to cluster this group without assigning revealed fraud to these clusters.

Cluster Score

Individually maximizing $C1$ and $C2$ leaves us in an unwanted situation. Basically, they are both trying to be split. However, notice that $\frac{x^j_{fraud}}{X^T_j} = \frac{1-x^j_{not-fraud}}{(X^T_j}$. Therefore, when we maximize one, we minimize the other. If we maximize both together, this results in a trade-off between the two, a trade-off in which we can choose. Moreover, as pointed out above, we actually want to maximize $C1$ subject to $C2$. Consequently, the fraud score is constructed as follows:

$$CS \;=\; (1+\alpha^2)\frac{C1 * C2}{C1 + C2 * \alpha}$$

With $\alpha 1$. If $\alpha = 1$, $C1$ and $C2$ will have the same weight. But if we assign $\alpha > 1$, this will reduce the charge of $C2$.
In conclusion, with this CS we have an objective parameter to tune the unsupervised model. Basically, we can maximize CS. The only decision that remains for us is to determine the relevance of α.

Practical Example

Imagine we have the following output from an unsupervised model:

Class	Label
0	1
0	2
0	3
0	1
1	2
1	2
1	3
0	3
0	3
0	2
0	1
0	3
1	2
0	1
1	2

Table 7. Class and Labels

The classes represent fraud (=1) and unlabeled (=0). The output label is the clustering label. As can be seen, just 33% of cases represent detected fraud. If we group the class by clusters:

Label	Class	Subtotal Class	Total
1	0	4	4
1	1	0	4
2	0	2	6
2	1	4	6
3	0	4	5
3	1	1	5

Table 8. Grouping labels and classes

As is evident, the fraud class tends to be assigned to the second cluster.
First we calculate $C1$.

$$C1 = \frac{\frac{0}{4} * 0 + \frac{4}{6} * 4 + \frac{1}{5} * 1}{5} = 0.5733$$

Then we calculate $C2$ using a similar formula.

$$C2 = \frac{\frac{4}{4} * 4 + \frac{2}{6} * 2 + \frac{4}{5} * 4}{10} = 0.7867$$

As can be seen, $C1$ gives worst results as its core group (group 2) is quite contaminated (66% of observations actually correspond to cases of fraud). This effect represents 93% of the total effect. The effect of mismatching the core group (1/5) is negligible, which stresses the importance of constructing a strong core group.
This conclusion is notorious in the case of C2. Non-identified classes are highly robust in two groups (1 and 3).
If we calculate the CS with $\alpha = 2$ (balanced $C1$ and $C2$) we obtain:

$$CS = 0.602$$

which is a value very close to 0.5733. This formula allows us to balance our results, giving greater weight to the lower score. We should stress we want both good and balanced scores; thus, $C1 = 0$, $C2 = 1$ is not the same as $C1 = 0.5$, $C2 = 0.5$. Indeed, the former returns a $CS = 0$. If we compare the mean with the CS:

$C1$	$C2$	**Mean**	CS
0.0	1.0	0.5	0.00
0.1	0.9	0.5	0.18
0.2	0.8	0.5	0.32
0.3	0.7	0.5	0.42
0.4	0.6	0.5	0.48
0.5	0.5	0.5	0.50
0.6	0.4	0.5	0.48
0.7	0.3	0.5	0.42
0.8	0.2	0.5	0.32
0.9	0.1	0.5	0.18
1.0	0.0	0.5	0.00

Table 9. Grouping labels and classes

As can be seen, we obtain the same unbalanced scores as the balanced outcomes for the mean score. CS penalizes the unbalanced scores. This is why we obtain different results with the same proportions.

However, we can make adjustments in terms of the relevance we attach to each group. If we raise α, we penalize the $C2$ results, and vice versa.

What happens if we choose $\alpha > 2$?

$C1$	$C2$	$\alpha = 2$	$\alpha = 4$	$\alpha = 6$
0.1	0.9	0.05	0.02	0.02
0.2	0.8	0.09	0.05	0.03
0.3	0.7	0.14	0.07	0.05
0.4	0.6	0.17	0.10	0.07
0.5	0.5	0.20	0.12	0.08
0.6	0.4	0.22	0.14	0.10
0.7	0.3	0.22	0.15	0.11
0.8	0.2	0.20	0.16	0.12
0.9	0.1	0.14	0.14	0.12

Table 10. Grouping labels and classes

As is evident, we obtain two effects. First, while $C1$ increases, CS also increases (although $C2$ decreases at the same rate). But the effect present in the balanced case now extends further. When we are at $C1 = 0.7$, the balanced effect tends to reverse the situation. The second effect is that the score curve is shifted downward. CS is now demanding more strongly that $C1$ be higher; while the higher the beta value the stronger $C1$.

Online Appendix 1. Data and Source Code

Data and source code can be founded here: `http://github.com/sebalp1987/outlier_model`

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2006

CREAP2006-01
Matas, A. (GEAP)**; Raymond, J.Ll.** (GEAP)
"Economic development and changes in car ownership patterns"
(Juny 2006)

CREAP2006-02
Trillas, F. (IEB)**; Montolio, D.** (IEB)**; Duch, N.** (IEB)
"Productive efficiency and regulatory reform: The case of Vehicle Inspection Services"
(Setembre 2006)

CREAP2006-03
Bel, G. (PPRE-IREA)**; Fageda, X.** (PPRE-IREA)
"Factors explaining local privatization: A meta-regression analysis"
(Octubre 2006)

CREAP2006-04
Fernàndez-Villadangos, L. (PPRE-IREA)
"Are two-part tariffs efficient when consumers plan ahead?: An empirical study"
(Octubre 2006)

CREAP2006-05
Artís, M. (AQR-IREA)**; Ramos, R.** (AQR-IREA)**; Suriñach, J.** (AQR-IREA)
"Job losses, outsourcing and relocation: Empirical evidence using microdata"
(Octubre 2006)

CREAP2006-06
Alcañiz, M. (RISC-IREA)**; Costa, A.; Guillén, M.** (RISC-IREA)**; Luna, C.; Rovira, C.**
"Calculation of the variance in surveys of the economic climate"
(Novembre 2006)

CREAP2006-07
Albalate, D. (PPRE-IREA)
"Lowering blood alcohol content levels to save lives: The European Experience"
(Desembre 2006)

CREAP2006-08
Garrido, A. (IEB)**; Arqué, P.** (IEB)
"The choice of banking firm: Are the interest rate a significant criteria?"
(Desembre 2006)

CREAP2006-09
Segarra, A. (GRIT)**; Teruel-Carrizosa, M.** (GRIT)
"Productivity growth and competition in spanish manufacturing firms:
What has happened in recent years?"
(Desembre 2006)

CREAP2006-10
Andonova, V.; Díaz-Serrano, Luis. (CREB)
"Political institutions and the development of telecommunications"
(Desembre 2006)

CREAP2006-11
Raymond, J.L.(GEAP)**; Roig, J.L..** (GEAP)
"Capital humano: un análisis comparativo Catalunya-España"
(Desembre 2006)

CREAP2006-12
Rodríguez, M.(CREB)**; Stoyanova, A.** (CREB)
"Changes in the demand for private medical insurance following a shift in tax incentives"
(Desembre 2006)

CREAP2006-13
Royuela, V. (AQR-IREA)**; Lambiri, D.**; **Biagi**, **B.**
"Economía urbana y calidad de vida. Una revisión del estado del conocimiento en España"
(Desembre 2006)

CREAP2006-14

Camarero, M.; Carrion-i-Silvestre, J.LL. (AQR-IREA).;**Tamarit**, **C.**
"New evidence of the real interest rate parity for OECD countries using panel unit root tests with breaks"
(Desembre 2006)

CREAP2006-15
Karanassou, M.; Sala, H. (GEAP).;**Snower** , **D. J.**
"The macroeconomics of the labor market: Three fundamental views"
(Desembre 2006)

2007

XREAP2007-01
Castany, L (AQR-IREA)**; López-Bazo, E.** (AQR-IREA).;**Moreno** , **R.** (AQR-IREA)
"Decomposing differences in total factor productivity across firm size"
(Març 2007)

XREAP2007-02
Raymond, J. Ll. (GEAP)**; Roig, J. Ll.** (GEAP)
"Una propuesta de evaluación de las externalidades de capital humano en la empresa"
(Abril 2007)

XREAP2007-03
Durán, J. M. (IEB)**; Esteller, A.** (IEB)
"An empirical analysis of wealth taxation: Equity vs. Tax compliance"
(Juny 2007)

XREAP2007-04
Matas, A. (GEAP)**; Raymond, J.Ll.** (GEAP)
"Cross-section data, disequilibrium situations and estimated coefficients: evidence from car ownership demand"
(Juny 2007)

XREAP2007-05
Jofre-Montseny, J. (IEB)**; Solé-Ollé, A.** (IEB)
"Tax differentials and agglomeration economies in intraregional firm location"
(Juny 2007)

XREAP2007-06
Álvarez-Albelo, C. (CREB)**; Hernández-Martín, R.**
"Explaining high economic growth in small tourism countries with a dynamic general equilibrium model"
(Juliol 2007)

XREAP2007-07
Duch, N. (IEB)**; Montolio, D.** (IEB); **Mediavilla, M.**
"Evaluating the impact of public subsidies on a firm's performance: a quasi-experimental approach"
(Juliol 2007)

XREAP2007-08
Segarra-Blasco, A. (GRIT)
"Innovation sources and productivity: a quantile regression analysis"
(Octubre 2007)

XREAP2007-09
Albalate, D. (PPRE-IREA)
"Shifting death to their Alternatives: The case of Toll Motorways"
(Octubre 2007)

XREAP2007-10
Segarra-Blasco, A. (GRIT); **Garcia-Quevedo, J.** (IEB); **Teruel-Carrizosa, M.** (GRIT)
"Barriers to innovation and public policy in catalonia"
(Novembre 2007)

XREAP2007-11
Bel, G. (PPRE-IREA); **Foote, J.**
"Comparison of recent toll road concession transactions in the United States and France"
(Novembre 2007)

XREAP2007-12
Segarra-Blasco, A. (GRIT);
"Innovation, R&D spillovers and productivity: the role of knowledge-intensive services"
(Novembre 2007)

XREAP2007-13
Bermúdez Morata, Ll. (RFA-IREA); **Guillén Estany, M.** (RFA-IREA), **Solé Auró, A**. (RFA-IREA)
"Impacto de la inmigración sobre la esperanza de vida en salud y en discapacidad de la población española"
(Novembre 2007)

XREAP2007-14
Calaeys, P. (AQR-IREA); **Ramos, R.** (AQR-IREA), **Suriñach, J**. (AQR-IREA)
"Fiscal sustainability across government tiers"
(Desembre 2007)

XREAP2007-15
Sánchez Hugalbe, A. (IEB)
"Influencia de la inmigración en la elección escolar"
(Desembre 2007)

2008

XREAP2008-01
Durán Weitkamp, C. (GRIT)**; Martín Bofarull, M.** (GRIT) ; **Pablo Martí, F.**
"Economic effects of road accessibility in the Pyrenees: User perspective"
(Gener 2008)

XREAP2008-02
Díaz-Serrano, L.; Stoyanova, A. P. (CREB)
"The Causal Relationship between Individual's Choice Behavior and Self-Reported Satisfaction: the Case of Residential Mobility in the EU"
(Març 2008)

XREAP2008-03
Matas, A. (GEAP)**; Raymond, J. L.** (GEAP)**; Roig, J. L.** (GEAP)
"Car ownership and access to jobs in Spain"
(Abril 2008)

XREAP2008-04
Bel, G. (PPRE-IREA) **; Fageda, X.** (PPRE-IREA)
"Privatization and competition in the delivery of local services: An empirical examination of the dual market hypothesis"
(Abril 2008)

XREAP2008-05
Matas, A. (GEAP); **Raymond, J. L.** (GEAP); **Roig, J. L.** (GEAP)
"**Job accessibility and employment probability**"
(Maig 2008)

XREAP2008-06
Basher, S. A.; Carrión, J. Ll. (AQR-IREA)
Deconstructing Shocks and Persistence in OECD Real Exchange Rates
(Juny 2008)

XREAP2008-07
Sanromá, E. (IEB)**; Ramos, R.** (AQR-IREA); Simón, H.
Portabilidad del capital humano y asimilación de los inmigrantes. Evidencia para España
(Juliol 2008)

XREAP2008-08
Basher, S. A.; Carrión, J. Ll. (AQR-IREA)
Price level convergence, purchasing power parity and multiple structural breaks: An application to US cities
(Juliol 2008)

XREAP2008-09
Bermúdez, Ll. (RFA-IREA)
A priori ratemaking using bivariate poisson regression models
(Juliol 2008)

XREAP2008-10
Solé-Ollé, A. (IEB), **Hortas Rico, M.** (IEB)
Does urban sprawl increase the costs of providing local public services? Evidence from Spanish municipalities
(Novembre 2008)

XREAP2008-11
Teruel-Carrizosa, M. (GRIT), **Segarra-Blasco, A.** (GRIT)
Immigration and Firm Growth: Evidence from Spanish cities
(Novembre 2008)

XREAP2008-12
Duch-Brown, N. (IEB), **García-Quevedo, J.** (IEB), **Montolio, D.** (IEB)
Assessing the assignation of public subsidies: Do the experts choose the most efficient R&D projects?
(Novembre 2008)

XREAP2008-13
Bilotkach, V., **Fageda, X.** (PPRE-IREA), **Flores-Fillol, R.**
Scheduled service versus personal transportation: the role of distance
(Desembre 2008)

XREAP2008-14
Albalate, D. (PPRE-IREA), **Gel, G.** (PPRE-IREA)
Tourism and urban transport: Holding demand pressure under supply constraints
(Desembre 2008)

2009

XREAP2009-01
Calonge, S. (CREB)**; Tejada, O.**
"A theoretical and practical study on linear reforms of dual taxes"
(Febrer 2009)

XREAP2009-02
Albalate, D. (PPRE-IREA)**; Fernández-Villadangos, L.** (PPRE-IREA)
"Exploring Determinants of Urban Motorcycle Accident Severity: The Case of Barcelona"
(Març 2009)

XREAP2009-03
Borrell, J. R. (PPRE-IREA)**; Fernández-Villadangos, L.** (PPRE-IREA)
"Assessing excess profits from different entry regulations"
(Abril 2009)

XREAP2009-04
Sanromá, E. (IEB)**; Ramos, R.** (AQR-IREA), **Simon, H.**
"Los salarios de los inmigrantes en el mercado de trabajo español. ¿Importa el origen del capital humano?"
(Abril 2009)

XREAP2009-05
Jiménez, J. L.; Perdiguero, J. (PPRE-IREA)
"(No)competition in the Spanish retailing gasoline market: a variance filter approach"
(Maig 2009)

XREAP2009-06
Álvarez-Albelo,C. D. (CREB)**, Manresa, A.** (CREB)**, Pigem-Vigo, M.** (CREB)
"International trade as the sole engine of growth for an economy"
(Juny 2009)

XREAP2009-07
Callejón, M. (PPRE-IREA)**, Ortún V, M.**
"The Black Box of Business Dynamics"
(Setembre 2009)

XREAP2009-08
Lucena, A. (CREB)
"The antecedents and innovation consequences of organizational search: empirical evidence for Spain"
(Octubre 2009)

XREAP2009-09
Domènech Campmajó, L. (PPRE-IREA)
"Competition between TV Platforms"
(Octubre 2009)

XREAP2009-10
Solé-Auró, A. (RFA-IREA)**,Guillén, M.** (RFA-IREA)**, Crimmins, E. M.**
"Health care utilization among immigrants and native-born populations in 11 European countries. Results from the Survey of Health, Ageing and Retirement in Europe"
(Octubre 2009)

XREAP2009-11
Segarra, A. (GRIT)**, Teruel, M.** (GRIT)
"Small firms, growth and financial constraints"
(Octubre 2009)

XREAP2009-12
Matas, A. (GEAP)**, Raymond, J.Ll.** (GEAP), **Ruiz, A.** (GEAP)
"Traffic forecasts under uncertainty and capacity constraints"
(Novembre 2009)

XREAP2009-13
Sole-Ollé, A. (IEB)
"Inter-regional redistribution through infrastructure investment: tactical or programmatic?"
(Novembre 2009)

XREAP2009-14
Del Barrio-Castro, T., **García-Quevedo, J.** (IEB)
"The determinants of university patenting: Do incentives matter?"
(Novembre 2009)

XREAP2009-15
Ramos, R. (AQR-IREA), **Suriñach, J.** (AQR-IREA)**, Artís, M.** (AQR-IREA)
"Human capital spillovers, productivity and regional convergence in Spain"
(Novembre 2009)

XREAP2009-16
Álvarez-Albelo, C. D. (CREB), **Hernández-Martín, R.**
"The commons and anti-commons problems in the tourism economy"
(Desembre 2009)

2010

XREAP2010-01
García-López, M. A. (GEAP)
"The Accessibility City. When Transport Infrastructure Matters in Urban Spatial Structure"
(Febrer 2010)

XREAP2010-02
García-Quevedo, J. (IEB), **Mas-Verdú, F.** (IEB), **Polo-Otero, J.** (IEB)
"Which firms want PhDs? The effect of the university-industry relationship on the PhD labour market"
(Març 2010)

XREAP2010-03
Pitt, D., **Guillén, M.** (RFA-IREA)
"An introduction to parametric and non-parametric models for bivariate positive insurance claim severity distributions"
(Març 2010)

XREAP2010-04
Bermúdez, Ll. (RFA-IREA), **Karlis, D.**
"Modelling dependence in a ratemaking procedure with multivariate Poisson regression models"
(Abril 2010)

XREAP2010-05
Di Paolo, A. (IEB)
"Parental education and family characteristics: educational opportunities across cohorts in Italy and Spain"
(Maig 2010)

XREAP2010-06
Simón, H. (IEB), **Ramos, R.** (AQR-IREA), **Sanromá, E.** (IEB)

"Movilidad ocupacional de los inmigrantes en una economía de bajas cualificaciones. El caso de España"
(Juny 2010)

XREAP2010-07
Di Paolo, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB)
"Language knowledge and earnings in Catalonia"
(Juliol 2010)

XREAP2010-08
Bolancé, C. (RFA-IREA), **Alemany, R.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"Prediction of the economic cost of individual long-term care in the Spanish population"
(Setembre 2010)

XREAP2010-09
Di Paolo, A. (GEAP & IEB)
"Knowledge of catalan, public/private sector choice and earnings: Evidence from a double sample selection model"
(Setembre 2010)

XREAP2010-10
Coad, A., **Segarra, A.** (GRIT), **Teruel, M.** (GRIT)
"Like milk or wine: Does firm performance improve with age?"
(Setembre 2010)

XREAP2010-11
Di Paolo, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB), **Calero, J.** (IEB)
"Exploring educational mobility in Europe"
(Octubre 2010)

XREAP2010-12
Borrell, A. (GiM-IREA), **Fernández-Villadangos, L.** (GiM-IREA)
"Clustering or scattering: the underlying reason for regulating distance among retail outlets"
(Desembre 2010)

XREAP2010-13
Di Paolo, A. (GEAP & IEB)
"School composition effects in Spain"
(Desembre 2010)

XREAP2010-14
Fageda, X. (GiM-IREA), **Flores-Fillol, R.**
"Technology, Business Models and Network Structure in the Airline Industry"
(Desembre 2010)

XREAP2010-15
Albalate, D. (GiM-IREA), **Bel, G.** (GiM-IREA), **Fageda, X.** (GiM-IREA)
"Is it Redistribution or Centralization? On the Determinants of Government Investment in Infrastructure"
(Desembre 2010)

XREAP2010-16
Oppedisano, V., **Turati, G.**
"What are the causes of educational inequalities and of their evolution over time in Europe? Evidence from PISA"
(Desembre 2010)

XREAP2010-17
Canova, L., **Vaglio, A.**
"Why do educated mothers matter? A model of parental help"
(Desembre 2010)

2011

XREAP2011-01
Fageda, X. (GiM-IREA), **Perdiguero, J.** (GiM-IREA)
"An empirical analysis of a merger between a network and low-cost airlines"
(Maig 2011)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

XREAP2011-02
Moreno-Torres, I. (ACCO, CRES & GiM-IREA)
"What if there was a stronger pharmaceutical price competition in Spain? When regulation has a similar effect to collusion"
(Maig 2011)

XREAP2011-03
Miguélez, E. (AQR-IREA); **Gómez-Miguélez, I.**
"Singling out individual inventors from patent data"
(Maig 2011)

XREAP2011-04
Moreno-Torres, I. (ACCO, CRES & GiM-IREA)
"Generic drugs in Spain: price competition vs. moral hazard"
(Maig 2011)

XREAP2011-05
Nieto, S. (AQR-IREA)**, Ramos, R.** (AQR-IREA)
"¿Afecta la sobreeducación de los padres al rendimiento académico de sus hijos?"
(Maig 2011)

XREAP2011-06
Pitt, D., **Guillén, M.** (RFA-IREA), **Bolancé, C.** (RFA-IREA)
"Estimation of Parametric and Nonparametric Models for Univariate Claim Severity Distributions - an approach using R"
(Juny 2011)

XREAP2011-07
Guillén, M. (RFA-IREA), **Comas-Herrera, A.**
"How much risk is mitigated by LTC Insurance? A case study of the public system in Spain"
(Juny 2011)

XREAP2011-08
Ayuso, M. (RFA-IREA), **Guillén, M.** (RFA-IREA), **Bolancé, C.** (RFA-IREA)
"Loss risk through fraud in car insurance"
(Juny 2011)

XREAP2011-09
Duch-Brown, N. (IEB), **García-Quevedo, J.** (IEB), **Montolio, D.** (IEB)
"The link between public support and private R&D effort: What is the optimal subsidy?"
(Juny 2011)

XREAP2011-10
Bermúdez, Ll. (RFA-IREA), **Karlis, D.**
"Mixture of bivariate Poisson regression models with an application to insurance"
(Juliol 2011)

XREAP2011-11
Varela-Irimia, X-L. (GRIT)
"Age effects, unobserved characteristics and hedonic price indexes: The Spanish car market in the 1990s"
(Agost 2011)

XREAP2011-12
Bermúdez, Ll. (RFA-IREA), **Ferri, A.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"A correlation sensitivity analysis of non-life underwriting risk in solvency capital requirement estimation"
(Setembre 2011)

XREAP2011-13
Guillén, M. (RFA-IREA), **Pérez-Marín, A.** (RFA-IREA), **Alcañiz, M.** (RFA-IREA)
"A logistic regression approach to estimating customer profit loss due to lapses in insurance"
(Octubre 2011)

XREAP2011-14
Jiménez, J. L., **Perdiguero, J.** (GiM-IREA), **García, C.**
"Evaluation of subsidies programs to sell green cars: Impact on prices, quantities and efficiency"
(Octubre 2011)

XREAP2011-15
Arespa, M. (CREB)
"A New Open Economy Macroeconomic Model with Endogenous Portfolio Diversification and Firms Entry"
(Octubre 2011)

XREAP2011-16
Matas, A. (GEAP), **Raymond, J. L.** (GEAP), **Roig, J.L.** (GEAP)
"The impact of agglomeration effects and accessibility on wages"
(Novembre 2011)

XREAP2011-17
Segarra, A. (GRIT)
"R&D cooperation between Spanish firms and scientific partners: what is the role of tertiary education?"
(Novembre 2011)

XREAP2011-18
García-Pérez, J. I.; **Hidalgo-Hidalgo, M.**; **Robles-Zurita, J. A.**
"Does grade retention affect achievement? Some evidence from PISA"
(Novembre 2011)

XREAP2011-19
Arespa, M. (CREB)
"Macroeconomics of extensive margins: a simple model"
(Novembre 2011)

XREAP2011-20
García-Quevedo, J. (IEB), **Pellegrino, G.** (IEB), **Vivarelli, M.**
"The determinants of YICs' R&D activity"
(Desembre 2011)

XREAP2011-21
González-Val, R. (IEB), **Olmo, J.**
"Growth in a Cross-Section of Cities: Location, Increasing Returns or Random Growth?"
(Desembre 2011)

XREAP2011-22
Gombau, V. (GRIT), **Segarra, A.** (GRIT)
"The Innovation and Imitation Dichotomy in Spanish firms: do absorptive capacity and the technological frontier matter?"
(Desembre 2011)

2012

XREAP2012-01
Borrell, J. R. (GiM-IREA), **Jiménez, J. L.**, **García, C.**
"Evaluating Antitrust Leniency Programs"
(Gener 2012)

XREAP2012-02
Ferri, A. (RFA-IREA), **Guillén, M.** (RFA-IREA), **Bermúdez, Ll.** (RFA-IREA)
"Solvency capital estimation and risk measures"
(Gener 2012)

XREAP2012-03
Ferri, A. (RFA-IREA), **Bermúdez, Ll.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"How to use the standard model with own data"
(Febrer 2012)

XREAP2012-04
Perdiguero, J. (GiM-IREA), **Borrell, J.R.** (GiM-IREA)
"Driving competition in local gasoline markets"
(Març 2012)

XREAP2012-05
D'Amico, G., **Guillen, M.** (RFA-IREA), Manca, R.
"Discrete time Non-homogeneous Semi-Markov Processes applied to Models for Disability Insurance"
(Març 2012)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

XREAP2012-06
Bové-Sans, M. A. (GRIT), Laguado-Ramírez, R.
"Quantitative analysis of image factors in a cultural heritage tourist destination"
(Abril 2012)

XREAP2012-07
Tello, C. (AQR-IREA), **Ramos, R.** (AQR-IREA), **Artís, M.** (AQR-IREA)
"Changes in wage structure in Mexico going beyond the mean: An analysis of differences in distribution, 1987-2008"
(Maig 2012)

XREAP2012-08
Jofre-Monseny, J. (IEB), **Marín-López, R.** (IEB), **Viladecans-Marsal, E.** (IEB)
"What underlies localization and urbanization economies? Evidence from the location of new firms"
(Maig 2012)

XREAP2012-09
Muñiz, I. (GEAP), **Calatayud, D.**, **Dobaño, R.**
"Los límites de la compacidad urbana como instrumento a favor de la sostenibilidad. La hipótesis de la compensación en Barcelona medida a través de la huella ecológica de la movilidad y la vivienda"
(Maig 2012)

XREAP2012-10
Arqué-Castells, P. (GEAP), **Mohnen, P.**
"Sunk costs, extensive R&D subsidies and permanent inducement effects"
(Maig 2012)

XREAP2012-11
Boj, E. (CREB), **Delicado, P.**, **Fortiana, J**., **Esteve, A.**, **Caballé, A.**
"Local Distance-Based Generalized Linear Models using the dbstats package for R"
(Maig 2012)

XREAP2012-12
Royuela, V. (AQR-IREA)
"What about people in European Regional Science?"
(Maig 2012)

XREAP2012-13
Osorio A. M. (RFA-IREA)**, Bolancé, C.** (RFA-IREA), **Madise, N.**
"Intermediary and structural determinants of early childhood health in Colombia: exploring the role of communities"
(Juny 2012)

XREAP2012-14
Miguelez. E. (AQR-IREA)**, Moreno, R.** (AQR-IREA)
"Do labour mobility and networks foster geographical knowledge diffusion? The case of European regions"
(Juliol 2012)

XREAP2012-15
Teixidó-Figueras, J. (GRIT)**, Duró, J. A.** (GRIT)
"Ecological Footprint Inequality: A methodological review and some results"
(Setembre 2012)

XREAP2012-16
Varela-Irimia, X-L. (GRIT)
"Profitability, uncertainty and multi-product firm product proliferation: The Spanish car industry"
(Setembre 2012)

XREAP2012-17
Duró, J. A. (GRIT), **Teixidó-Figueras, J.** (GRIT)
"Ecological Footprint Inequality across countries: the role of environment intensity, income and interaction effects"
(Octubre 2012)

XREAP2012-18
Manresa, A. (CREB), **Sancho, F.**
"Leontief versus Ghosh: two faces of the same coin"
(Octubre 2012)

XREAP2012-19
Alemany, R. (RFA-IREA), **Bolancé, C.** (RFA-IREA), **Guillén, M.** (RFA-IREA)
"Nonparametric estimation of Value-at-Risk"
(Octubre 2012)

XREAP2012-20
Herrera-Idárraga, P. (AQR-IREA), **López-Bazo, E.** (AQR-IREA), **Motellón, E.** (AQR-IREA)
"Informality and overeducation in the labor market of a developing country"
(Novembre 2012)

XREAP2012-21
Di Paolo, A. (AQR-IREA)
"(Endogenous) occupational choices and job satisfaction among recent PhD recipients: evidence from Catalonia"
(Desembre 2012)

2013

XREAP2013-01
Segarra, A. (GRIT), **García-Quevedo, J.** (IEB), **Teruel, M.** (GRIT)
"Financial constraints and the failure of innovation projects"
(Març 2013)

XREAP2013-02
Osorio, A. M. (RFA-IREA), **Bolancé, C.** (RFA-IREA), Madise, N., Rathmann, K.
"Social Determinants of Child Health in Colombia: Can Community Education Moderate the Effect of Family Characteristics?"
(Març 2013)

XREAP2013-03
Teixidó-Figueras, J. (GRIT), **Duró, J. A.** (GRIT)
"The building blocks of international ecological footprint inequality: a regression-based decomposition"
(Abril 2013)

XREAP2013-04
Salcedo-Sanz, S., **Carro-Calvo, L.**, **Claramunt, M.** (CREB), **Castañer, A.** (CREB), **Marmol, M.** (CREB)
"An Analysis of Black-box Optimization Problems in Reinsurance: Evolutionary-based Approaches"
(Maig 2013)

XREAP2013-05
Alcañiz, M. (RFA), **Guillén, M.** (RFA), **Sánchez-Moscona, D.** (RFA), **Santolino, M.** (RFA), **Llatje, O., Ramon, Ll.**
"Prevalence of alcohol-impaired drivers based on random breath tests in a roadside survey"
(Juliol 2013)

XREAP2013-06
Matas, A. (GEAP & IEB), **Raymond, J. Ll.** (GEAP & IEB), **Roig, J. L.** (GEAP)
"How market access shapes human capital investment in a peripheral country"
(Octubre 2013)

XREAP2013-07
Di Paolo, A. (AQR-IREA), **Tansel, A.**
"Returns to Foreign Language Skills in a Developing Country: The Case of Turkey"
(Novembre 2013)

XREAP2013-08
Fernández Gual, V. (GRIT), **Segarra, A.** (GRIT)
"The Impact of Cooperation on R&D, Innovation andProductivity: an Analysis of Spanish Manufacturing and Services Firms"
(Novembre 2013)

XREAP2013-09
Bahraoui, Z. (RFA); **Bolancé, C.** (RFA); **Pérez-Marín. A. M.** (RFA)
"Testing extreme value copulas to estimate the quantile"
(Novembre 2013)

2014

XREAP2014-01
Solé-Auró, A. (RFA), **Alcañiz, M.** (RFA)
"Are we living longer but less healthy? Trends in mortality and morbidity in Catalonia (Spain), 1994-2011"
(Gener 2014)

XREAP2014-02

Teixidó-Figueres, J. (GRIT), **Duro, J. A.** (GRIT)
"Spatial Polarization of the Ecological Footprint distribution"
(Febrer 2014)

XREAP2014-03
Cristobal-Cebolla, A.; Gil Lafuente, A. M. (RFA), **Merigó Lindhal, J. M.** (RFA)
"La importancia del control de los costes de la no-calidad en la empresa"
(Febrer 2014)

XREAP2014-04
Castañer, A. (CREB); **Claramunt, M.M.** (CREB)
"Optimal stop-loss reinsurance: a dependence analysis"
(Abril 2014)

XREAP2014-05
Di Paolo, A. (AQR-IREA); **Matas, A.** (GEAP); **Raymond, J. Ll.** (GEAP)
"Job accessibility, employment and job-education mismatch in the metropolitan area of Barcelona"
(Maig 2014)

XREAP2014-06
Di Paolo, A. (AQR-IREA); **Mañé, F.**
"Are we wasting our talent? Overqualification and overskilling among PhD graduates"
(Juny 2014)

XREAP2014-07
Segarra, A. (GRIT); **Teruel, M.** (GRIT); **Bové, M. A.** (GRIT)
"A territorial approach to R&D subsidies: Empirical evidence for Catalonian firms"
(Setembre 2014)

XREAP2014-08
Ramos, R. (AQR-IREA); **Sanromá, E.** (IEB); **Simón, H.**
"Public-private sector wage differentials by type of contract: evidence from Spain"
(Octubre 2014)

XREAP2014-09
Bel, G. (GiM-IREA); **Bolancé, C.** (Riskcenter-IREA); **Guillén, M.** (Riskcenter-IREA)**; Rosell, J.** (GiM-IREA)
"The environmental effects of changing speed limits: a quantile regression approach"
(Desembre 2014)

2015

XREAP2015-01
Bolance, C. (Riskcenter-IREA); **Bahraoui, Z.** (Riskcenter-IREA), **Alemany, R.** (Risckcenter-IREA)
"Estimating extreme value cumulative distribution functions using bias-corrected kernel approaches"
(Gener 2015)

XREAP2015-02
Ramos, R. (AQR-IREA); **Sanromá, E.** (IEB), **Simón, H.**
"An analysis of wage differentials between full- and part-time workers in Spain"
(Agost 2015)

XREAP2015-03
Cappellari, L.; **Di Paolo, A.** (AQR-IREA)
"Bilingual Schooling and Earnings: Evidence from a Language-in-Education Reform"
(Setembre 2015)

XREAP2015-04
Álvarez-Albelo, C. D., Manresa, A. (CREB)**, Pigem-Vigo, M.** (CREB)
"Growing through trade: The role of foreign growth and domestic tariffs"
(Novembre 2015)

XREAP2015-05
Caminal, R., Di Paolo, A. (AQR-IREA)
Your language or mine?
(Novembre 2015)

XREAP2015-06
Choi, H. (AQR-IREA)**, Choi, A.** (IEB)
When one door closes: the impact of the hagwon curfew on the consumption of private tutoring in the Republic of Korea

(Novembre 2015)

2016

XREAP2016-01
Castañer, A. (CREB, XREAP); **Claramunt, M M.** (CREB, XREAP), **Tadeo, A.**, **Varea, J.** (CREB, XREAP)
Modelización de la dependencia del número de siniestros. Aplicación a Solvencia II
(Setembre 2016)

XREAP2016-02
García-Quevedo, J. (IEB, XREAP); **Segarra-Blasco, A.** (GRIT, XREAP), **Teruel, M.** (GRIT, XREAP)
Financial constraints and the failure of innovation projects
(Setembre 2016)

XREAP2016-03
Jové-Llopis, E. (GRIT, XREAP); **Segarra-Blasco, A.** (GRIT, XREAP)
What is the role of innovation strategies? Evidence from Spanish firms
(Setembre 2016)

XREAP2016-04
Albalate, D. (GiM-IREA, XREAP); **Rosell, J.** (GiM-IREA, XREAP)
Persistent and transient efficiency on the stochastic production and cost frontiers – an application to the motorway sector
(Octubre 2016)

XREAP2016-05
Jofre-Monseny, J. (IEB, XREAP), **Silva, J. I.**, **Vázquez-Grenno, J.** (IEB, XREAP)
Local labor market effects of public employment
(Novembre 2016)

XREAP2016-06
Garcia-López, M. A. (IEB, XREAP), **Hemet, C.**, **Viladecans-Marsal, E.** (IEB, XREAP)
Next train to the polycentric city: The effect of railroads on subcenter formation
(Novembre 2016)

XREAP2016-07
Vayá, E. (AQR-IREA, XREAP)**, García, J. R.** (AQR-IREA, XREAP)**, Murillo, J.** (AQR-IREA, XREAP)**, Romaní, J.** (AQR-IREA, XREAP)**, Suriñach, J.** (AQR-IREA, XREAP)**,**
Economic impact of cruise activity: the port of Barcelona
(Desembre 2016)

XREAP2016-08
Ayuso, M. (Riskcenter, XREAP)**, Guillen, M.** (Riskcenter, XREAP)**, Nielsen, J. P.**
Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data
(Desembre 2016)

XREAP2016-09
Ruíz, A. (GEAP, XREAP), **Matas, A.** (GEAP, XREAP), **Raymond, J. Ll.**
How do road infrastructure investments affect the regional economy? Evidence from Spain
(Desembre 2016)

2017

XREAP2017-01
Bernardo, V. (GiM-IREA, XREAP); **Fageda, X.** (GiM-IREA, XREAP)
Globalization, long-haul flights and inter-city connections
(Octubre 2017)

XREAP2017-02
Di Paolo, A. (AQR-IREA, XREAP); **Tansel, A.**
Analyzing Wage Differentials by Fields of Study: Evidence from Turkey
(Octubre 2017)

XREAP2017-03
Melguizo, C. (AQR-IREA, XREAP); **Royuela, V.** (AQR-IREA, XREAP)
What drives migration moves across urban areas in Spain? Evidence from the great recession
(Octubre 2017)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

XREAP2017-04
Boonen, T.J., Guillén, M. (RISKCENTER, XREAP); **Santolino, M.** (RISKCENTER, XREAP)
Forecasting compositional risk allocations
(Octubre 2017)

XREAP2017-05
Curto-Grau, M. (IEB, XREAP)**, Solé-Ollé, A.** (IEB, XREAP)**, Sorribas-Navarro, P.** (IEB, XREAP)
Does electoral competition curb party favoritism?
(Novembre 2017)

XREAP2017-06
Esteller, A. (IEB, XREAP)**, Piolatto, A.** (IEB, XREAP)**, Rablen, M. D.**
Taxing high-income earners: tax avoidance and mobility
(Novembre 2017)

XREAP2017-07
Bolancé, C. (RISKCENTER, XREAP), **Vernic, R**
Multivariate count data generalized linear models: Three approaches based on the Sarmanov distribution
(Novembre 2017)

XREAP2017-08
Albalate, D. (GiM-IREA, XREAP)**, Bel-Piñana, P. (**GiM-IREA, XREAP)
Public Private Partnership management effects on road safety outcomes
(Novembre 2017)

XREAP2017-09
Teruel, M. (GRIT, XREAP)**, Segarra, A. (G**RIT, XREAP)
Gender diversity, R&D teams and patents: An application to Spanish firms
(Novembre 2017)

XREAP2017-10
Cuberes, D., **Teignier, M.** (CREB, XREAP)
How Costly Are Labor Gender Gaps? Estimates by Age Group for the Balkans and Turkey
(Novembre 2017)

XREAP2017-11
Murilló, I. P., **Raymond, J. L.** (GEAP, XREAP), **Calero, J.** (IEB, XREAP)
Efficiency in the transformation of schooling into competences: A cross-country analysis using PIAAC data
(Novembre 2017)

XREAP2017-12
Giuntella, O., Mazzonnay, F., Nicodemo, C. (GEAP, XREAP), Vargas Silva, C
Immigration and the Reallocation of Work Health Risks
(Desembre 2017)

XREAP2017-13
Giuntella, O., Nicodemo, C. (GEAP, XREAP), Vargas Silva, C.
The Effects of Immigration on NHS Waiting Times
(Desembre 2017)

XREAP2017-14
Solé-Ollé, A. (IEB, XREAP), Viladecans-Marsal, E. (IEB, XREAP)
Housing Booms and Busts and Local Fiscal Policy
(Desembre 2017)

XREAP2017-15
Jové-Llopis, E. (GRIT, XREAP), Segarra-Blasco, A. (GRIT, XREAP)
Eco-strategies and firm growth in European SMEs
(Desembre 2017)

SÈRIE DE DOCUMENTS DE TREBALL DE LA XREAP

2018

XREAP2018-01
Teruel, E. (GRIT, XREAP), Segarra-Blasco, A. (GRIT, XREAP)
Gender diversity, R&D teams and patents: An application to Spanish firms
(Febrer 2018)

XREAP2018-02
Palacio, S. M. (GiM, XREAP)
Detecting Outliers with Semi-Supervised Machine Learning: A Fraud Prediction Application
(Abril 2018)

xarxa.xreap@gmail.com