COMPLUTENSE UNIVERSITY MADRID

FACULTY OF MATHEMATICAL SCIENCE

MASTER IN OFFICIAL STATISTICS AND SOCIAL AND ECONOMIC INDICATORS

MASTER THESIS

2019-2020

Selective Data Editing of Continuous Variables with Random Forests in Official Statistics

---

Author:
Sarah BOHNENSTEFFEN

Supervisor:
Dr. David SALGADO FERNANDEZ
Elena ROSA PEREZ

September 6, 2020

Declaration of Authorship

I, Sarah BOHNENSTEFFEN, declare that this thesis titled, "Selective Data Editing of Continuous Variables with Random Forests in Official Statistics" and the work presented in it are my own. I confirm that:

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: BOHNENSTEFFEN SARAH - Y6944875Q Digitally signed by BOHNENSTEFFEN SARAH - Y6944875Q Date: 2020.09.06 22:51:32 +02'00'

Date: *September 6, 2020*

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."

John Tukey

COMPLUTENSE UNIVERSITY MADRID

Abstract

Faculty of Mathematical Science

European Master in Official Statistics

Selective Data Editing of Continuous Variables with Random Forests in Official Statistics

by Sarah BOHNENSTEFFEN

Technological advances and new demands due to economic and socio-cultural changes regularly challenge the National Statistical Institutes to adapt to their evolving environment. The application of machine learning methods as important and promising tools for official statistics are discussed in the context of these changes, in the context of opportunities arising from new digital data sources, and considering the difficult task of having to balance a variety of quality requirements at national and international level. Selective statistical data editing is an approach to detect influential units and select them for manual follow up in order to make the process more efficient. In this thesis, a simple and a two-step approach are developed to apply random forests to selective editing of continuous variables in the context of short-term business survey data. We present a score function based on decision forest models which allows for an efficient selection of units relevant for the estimation of the final estimates. The approach is found to be applicable also at the disaggregated levels of the autonomous communities and economic branches.

El avance tecnológico y nuevas demandas debidas a cambios económicos y socioculturales desafían regularmente a los Institutos Nacionales de Estadística a adaptarse a su entorno en constante evolución. La aplicación de métodos de aprendizaje automático como instrumentos importantes y prometedores para las estadísticas oficiales se analizan en el contexto de esos cambios, en el contexto de las oportunidades que surgen de nuevas fuentes de datos digitales, y teniendo en cuenta la difícil tarea de tener que equilibrar una variedad de requisitos de calidad a nivel nacional e internacional. La depuración selectiva es un conjunto de técnicas para detectar unidades influyentes y seleccionarlas para el seguimiento manual a fin de hacer el proceso más eficiente. En este trabajo se desarrolla un enfoque simple y uno en dos etapas para aplicar los bosques aleatorios a la depuración selectiva de variables continuas en el contexto de datos de encuestas económicas coyunturales. Se presenta una función de puntuación basada en modelos de bosques aleatorios que permite una selección eficiente de unidades relevantes para la estimación de los agregados finales. El enfoque desarrollado también es aplicable a los niveles desagregados de las comunidades autónomas y ramas de negocio para los datos usados.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The application of machine learning solutions in official statistics has become an extremely prominent topic during the last years. Just recently, the UNECE Statistical Data Editing Virtual Workshop 2020, held in early September, was dominated by topics like machine learning and artificial intelligence and how to use these methods in the context of the official data production process.

Machine learning is not exactly a new topic. The fast expansion of computer power and the improvements in information infrastructure development in the last half-century have made the implementation of existing data analysis techniques possible and have provided a breeding ground for the development of many other methods. Coupled with the ever-increasing volume of available data and the development of more complex forms and data formats, especially unstructured data, techniques such as support vector machines, neural networks, random forests and others have become increasingly popular in recent decades (Biamonte et al., 2017). However, National Statistics Instituts (NSIs) must overcome particular challenges when developing new methodologies and applying new methods. As a legally legitimized bureaucratic organization, their processes are subject to certain rules and standards. Their data are an important basis for decision-making for political, economic and social actors in democratic societies. As part of the European Statistical System, the NSIs are committed to quality and stand for the accuracy and reliability of the information they provide (MacFeely, 2016).

On the other hand, the NSIs are also called upon to provide relevant and up-to-date information and, in addition, to improve timeliness and punctuality of their production and dissemination practice. This puts them in the difficult situation of competing with non-public data providers on the one hand, and on the other, having to manage the conflict that emerges between these goals and further requirements such as cost efficiency of their work and to maintain their professional independence despite the dependence on public funds (Ljones, 2011; Sæbø and Holmberg, 2019). Thus, this thesis focuses on a point where a huge potential for optimization is identified, namely the reduction of the necessary manual resources through a more efficient data editing process. This could also improve the quality of the data, either directly through the increased coherence of the process or indirectly by reallocating manual resources to areas where the data production process can be improved otherwise.

This work was developed in the context of an internship at the Methodological Department at National Statistical Institute of Spain (Statistics Spain) under the supervision of Dr. David Salgado. The unit develops new statistical methodologies for a standardized official statistical production, as well as unified software tools for their implementation. The questions that this thesis is based on are *how the official statistical data editing process can be improved by using machine learning methods*, namely random forests, and *which challenges are associated with the application of such methods*

in the context of the official statistical production. As part of data processing, editing is a crucial and resource-consuming part of the work of the national statistical offices (Arbues et al., 2013). Erroneous values in the raw data must be detected and treated so that the data meet quality standards when they are further processed in the estimation phase. Especially in business surveys, editing and imputation is identified as one of the most resource and time-consuming survey processes. Over the years, the statistical offices have learned from their experience that particularly influential observations can be identified which are responsible for much of the inaccuracy in final estimates (Scholtus et al., 2014). Techniques ranking observations according to their potential influential errors in order to select units for further interactive editing are jointly known as selective editing. To develop a random forest model that is be able to compute the score function on which selective editing is based, data from the Short-term Business Survey SSAI conducted by Statistics Spain will be used. The idea is to compute the error between the raw turnover value which was reported by the units and the edited turnover value after it was manually revised. This error combined with the sample weight of each unit can then be used to predict which units should be given preference in further manual editing steps. Random forests seem to be a convincing method for this, as they can be used with both numerical and categorical predictors for both regression or classification tasks. Random forests can handle non-linear features and they are robust to the inclusion of irrelevant predictors, they are parallelizable and scalable (Cutler et al., 2011).

To address the questions and presented issues this thesis is structured as follows: chapter 2 outlines the theoretical background of data editing in Official Statistics. It discusses the position of NSIs between national and European responsibility, explains the quality requirements for official statistics and the resulting challenges and introduces selective editing. Finally, the selective editing approaches used in this thesis are presented. Chapter 3 introduces the Services Sector Activity Indicators as data basis and presents characteristics of short-term business survey data. Chapter 4 gives a short introduction to what machine learning is, provides an overview of the current state of research of machine learning in official statistics, explains the concept of random forests and gives some notes on the software used. In chapter 5 the practical steps and the respective results of both approaches are explained. Our investigation was originally intended to analyse the application of random forests to continuous variables such as the described error variable. It turned out, however, that the chosen target variable is rather a semi-continuous variable, so a second, two-step approach was developed to adapt the procedure to the properties of the data and to do justice to its characteristics. Chapter 5 explains the practical steps and the respective results of these two approaches, before the final chapter summarizes the work and its results, discusses open points and presents remaining questions.

Chapter 2

Theoretical Framework: Data Editing in the context of National Statistical Institutes

2.1 Quality Requirements in Official Statistics

The work of the NSIs is exposed to constant changes in external circumstances. Technical innovations, the emergence of new data types and dimensions (Big Data, social network data) and changing socioeconomic realities shaped by globalization. In the last years, NSI have dealt with a lot of important questions like how to make use of Big Data, how to reduce the response burden for firms and individuals or how to assess quality of non-survey data sources (MacFeely, 2016).

A constantly changing environment is also a challenge for other institutions and companies. Nevertheless, National Statistical Institutes face particular challenges due to their quality standards, bureaucratic organizational structure and legal responsibilities. In the following sections, we will therefore start by discussing the role and working environment of Statistics Spain (which can largely be generalized to other NSIs), in order to gain an understanding of the relevance of an efficient data editing process.

2.1.1 National Statistical Institutes as bureaucratic organizations

Similarly to statistical systems in other countries, the Spanish National Statistical System is responsible for the compilation of both official statistics for state purposes and European statistics. Analyzing the organizational set-up of the structures of Statistics Spain, many similarities can be found with the prototype called *bureaucratic organization* as described by Weber (1978, p. 217ff.):

In accordance with the principle of administrative records, all types of administrative acts, decisions and instructions are recorded in writing. Bureaucratic organizations also keep the principle of separating resources and staff: The administrative staff does not own the material means or resources of the administration. Furthermore, each office has its own specific sphere of competence. Offices are organized in a hierarchical way, meaning that each lower office is under the control and supervision of a higher one. At Statistics Spain, this principle is applied at the geographical level in the provincial delegations that supply the central services, as well as within the central office, in which the general subdirectorates are subordinate to the departments, and which again are divided into different work areas.

The administrative staff is selected on the basis of the technical qualifications of the candidates, which are "tested by examination or guarantied by diplomas certifying technical training, or both" (Weber, 1978, p. 220), which in case of Statistics Spain

corresponds to the implementation of a civil service entrance examination (*oposiciones*) with minimum qualifications.

The legitimacy of a bureaucratic organization is based on constitutional principles and the rule of law (Olsen, 2008). As part of the national statistical system, the National Statistical Institute of Spain (Statistics Spain) has the goal and the legal mandate to provide social and political actors with high quality statistical information. These tasks and responsibilities are assigned to Statistics Spain within the framework of specific regulations and legal norms. Statistical activity for State purposes is anchored in the Spanish Constitution (149.1.31., 1978) as an exclusive competence of the State. Statistics Spain is a legally independent administrative autonomous institution assigned to the Ministry of Economic Affairs and Digital Transformation, via the Secretary of State for the Economy and Business Support. The legal basis of the work of Statistics Spain is the Law 12/1989 of 9 May 1989 on the Public Statistical Services (LFEP).

According to bureaucratic organizational theory, bureaucratic organizations enjoy technical superiority over other forms of organization and are designed to ensure efficiency and economic effectiveness (Kim et al., 2014). Nevertheless, this form of organization has been subject to sharp criticism in the 70s and ever since. The term in its negative meaning refers to a form of organization characterized by slowness and inefficiency. Contrary to Weber's idea, who described bureaucratic organization as the most modern and efficient form of organization, numerous studies on bureaucracy in public research institutions suggest that these are inefficient because of their public nature, which is associated with red tape, the negative by-product of bureaucracy (e.g. Coccia, 2009; Crow and Bozeman, 1989). Bureaucracy seems to be an obstacle to keeping up with market-oriented companies and is accused of hindering innovation and devouring resources. A particularly dominant argument is that bureaucratically organized processes are lengthy and therefore do not offer the possibility of reacting appropriately quickly and flexibly to changes.

On the other hand, the under-complexity of "bureaucracy bashing" (Olsen, 2008) and the lack of a clear definition of the term has been criticized and shifting the focus from seeing difficulties to seeking chances, it has been argued bureaucracy and efficiency or technology should not be seen as antagonists, but as complementary. Taking standardization that comes with bureaucracy as an example: the effort required to develop a standard suitable for all procedures in a given scope can seem time consuming and burdensome. Things change if it is seen as part of the solution. Combined with international cooperation, standardization can even increase the flexibility of the institution, as well as its ability to respond to change, because instruments developed e.g. by other NSIs may be easily adaptable to and compatible with existing structures. For example, the Common Statistical Production Architecture (CSPA), a standard to share common functionalities inside the European Statistical System, is based on this argument (Vale, 2014).

A study of the interaction relationship between bureaucracy and information technology, suggests that the use of IT, which led to improved decision-making, could reduce bureaucracy and sectarianism in organizations in the long run. The study also concludes that the technical competence of staff is positively correlated with the time saved through the increased use of IT systems (Kim et al., 2014). The idea of mechanical rationalization as a common goal of bureaucracy and IT agrees with the Weberian image of the bureaucratic organization as efficient through technical competence and the above mentioned results suggest that technical innovation could be a key element enabling bureaucracy to counteract the problems it inherently faces. Without going into further detail, the relationships described will be

kept in mind in the following remarks.

2.1.2 Quality Dimensions (Im-)balance

To understand the working environment of NSIs beyond their organisational structure, it is essential to understand the complex situation of their areas of responsibility. The NSIs have a shared competence and responsibility between their governmental mandate in the national context on the one hand and their integration into the European statistical network on the other. It is generally agreed that public research institutions play an essential role in modern economies and are also an integral institution for a functioning democracy, which is why independent data of the highest quality are of enormous importance (MacFeely, 2016, p. 789). With the transition to a data-driven management of public administrations, the dependence on high-quality data is increasing and becomes crucial to invest public money effectively (Ljones, 2011, p. 27). The role of NSIs as European statistical producers gives rise to diverse quality requirements. While the emphasis on quality has always been an important characteristic of official statistics, quality dimensions have been extended and made more specific in the last years (Ljones, 2011, p. 25).

A classic approach to detect statistical error sources and properties in the context of quality assurance is the Total Survey Error (TSE) Framework (Biemer, 2010; Groves and Lyberg, 2010), which divides error sources into the two main dimensions *measurement* and *representation*. Following the terminology of the TSE Framework, the problem addressed in this thesis is situated on the measurement side between *measurement* and *edited response*, where either a measurement error or a processing error can occur. Both are not necessarily detected in the editing process; the true value always remains unknown. Nevertheless, there may be errors that occur systematically or at least with a certain frequency in relation to other known variables. The approach can therefore be seen as an attempt to learn from predictable errors in this part, so that editing resources can be efficiently directed to where we can most likely expect these errors. The TSE framework has also been the basis to make it broader by including errors arising from the use (and integration) of different data sources like administrative data.

Nevertheless, such a perspective would not be sufficient to describe the comprehensive quality requirements for the work of National Statistical Institutes like INE. As member state of the European Union, Spains National Statistical Institute is part of the European Statistical System (ESS). The ESS is a partnership between Eurostat and the statistical authorities at the national level, which is responsible for the development, production and dissemination of European statistics. The legal basis of european national statistical institutes like Statistics Spain is the Regulation (EC) No. 223/2009 on European Statistics (EU, 2009), which is derived from the Treaty of the Functioning of the European Union.

The two pillars of quality assurance in the European Statistical System are the European Statistics Code of Practice[1] (Eurostat, 2017) and the Quality Assurance Framework of the European Statistical System[2] (ESS, 2019). The ES CoP in its latest version from 2017 is an update of the first one originally adopted in 2005. It is a self-regulatory tool, meaning that its assessment is done by the institutions that implement it. In part, it coincides with the guidelines laid down in the European Statistical Law, such as professional independence, the coordination role of Eurostat and the NSIs statistical confidentiality, but it also complements them. In its current

[1] From here on referred to as ES CoP.

[2] From here on referred to as ESS QAF

form, it consists of the following 16 principles. Principles 1 to 6 refer to the institutional environment, while principles 7 to 10 refer to good practices regarding the statistical processes and principles 11 to 15 are concerning the statistical output.

	Principle	
1.	**Professional Independence**	Professional independence means independence of statistical authorities from other policy, regulatory or administrative departments and bodies, as well as from private sector operators.
1b.	**Coordination and cooperation**	The coordination of all activities for the development, production and dissemination of European statistics is ensured by the National Statistical Institutes at the level of the national statistical system and by Eurostat at the level of the European Statistical System.
2.	**Mandate for Data Collection and Access to Data**	Statistical authorities have a legal mandate to collect and access information from multiple data sources, like data from administrations, enterprises or households, for European statistical purposes.
3.	**Adequacy of Resources**	The human, financial and technical resources available to statistical authorities are sufficient regarding both their magnitude and quality to meet European Statistics requirements.
4.	**Commitment to Quality**	Statistical authorities are committed to quality and an organizational structure and procedures are in place to manage, monitor and improve quality management.
5.	**Statistical Confidentiality and Data Protection**	The privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and its use only for statistical purposes are absolutely guaranteed.
6.	**Impartiality and Objectivity**	Statistical authorities develop, produce and disseminate European Statistics respecting scientific independence and in an objective, professional and transparent manner, in which all users are treated equitably.
7.	**Sound Methodology**	The overall methodological framework used for European Statistics follows European and other international guidelines, and good practices. It is based on the application of standards and adequate tools, constantly striving for innovation. Continuous training for their staff is implemented.
8.	**Appropriate Statistical Procedures**	The implementation of appropropiate statistical procedures throughout the statistical processes is ensured by making use of definitions and concepts of data for non-statistical purposes, pretesting questionnaires, establishing cooperations and agreements with data holders, the collection of metadata and transparent revisions.
9.	**Non-excessive Burden on Respondents**	The response burden is proportionate to the needs of the users and is not excessive for respondents. European Statistics demands are limited to the absolutely necessary and the statistical authorities monitor the response burden and set targets for its reduction. Administrative and other data sources are used whenever possible.

10. **Cost Effectiveness**	Available Resources are used effectively by optimizing the use of information and communication technology, improving the statistical potential of administrative and other data sources and the implementation of standardized solutions. The use of resources is internally and externally monitored.
11. **Relevance**	European Statistics are based on the needs of users. For this end, procedures are in place to identify users needs, consult users, to monitor the relevance of existing statistics and monitor user satisfaction.
12. **Accuracy and Reliability**	Source data, integrated data, intermediate results and statistical outputs are regularly assessed and validated to ensure that European Statistics accurately and reliably portray reality.
13. **Timeliness and Punctuality**	Statistics are released in a timely and punctual manner, at a standard daily time, following the release calendar. Divergence from the dissemination time schedule is publicised in advance, explained and a new release date set.
14. **Coherence and Comparability**	Statistics are coherent and consistent internally, over time and comparable between regions and countries. Statistics from the different data sources and of different periodicity are compared and reconciled. Cross-national comparability of the data is ensured by periodical exchanges between the ESS and other statistical systems.
15. **Accessibility and Clarity**	Statistics are presented in a clear and understandable form, released in a suitable and convenient manner, available and accessible on an impartial basis with supporting metadata, using modern information and communication technology and open data standards.

TABLE 2.2: Key Points of the European Statistics Code of Practice Principles (Eurostat, 2017)

Although a typical association of quality measures in the context of data would be statistical concepts like accuracy, looking at the principles outlined above, one can see that quality is a much wider concept in the context of official statistics. Moreover, it is constantly reviewed, enhanced and made more concrete. This is why the ES CoP is periodically updated since its adaption in 2005. Due to their comprehensive nature the ES CoP principles remain rather superficial and of general nature, which may make them difficult to apply and assess (Eurostat, 2017; Revilla and Piñán, 2012). In order to facilitate the implementation of the ES CoP and to give concepts a more concrete form, the ESS QAF offers a collection of methods and tools, giving examples of good practices.

A further problem related to this extensive collection of principles is their multidimensionality. Although the ESS QAF provides guidance for its implementation, concepts of different quality dimensions may appear contradictory in practical application. This makes them difficult to apply concurrently. Saebo (2019) points out that there are in fact trade-offs between different principles of the quality framework. In identifying conflicting pairs of principles, it is noted that compliance with the various principles must be carefully balanced. Professional independence and impartiality may be challenged by the principle of cooperation, for example, or may

conflict with the principle of relevance. It is immediately apparent that principles of accuracy (P12), timeliness (P13) and cost-effectiveness (P10) are competing concepts (Sæbø and Holmberg, 2019).

Furthermore, difficulties in adhering to the concept of cost-effectiveness intensify problems of contradictions in the relationships between other principles. The greater the scarcity of resources, the greater the difficulty in compliance with other principles. This leads to structural dependencies on the government, in a direct way from public budgets or in an indirect way due to the need of less cost-intensive data sources, like the administrative data, the obtaining of which, however, is often associated with a huge bureaucracy effort, and whose concepts and collection methods are again based on the government's design. Surveys collected by NSIs themselves have the advantage that the NSIs can decide on how the data are generated (e.g. decisions about the survey format, components of the questionnaire or the survey cycle). Conversely, this competence goes hand in hand with the responsibility to subject all phases of data generation to quality analysis and to actively use the metadata of the production process to optimize procedures. In addition to cost effectiveness, there are challenges such as keeping up with other statistics producers in terms of timeliness (P13) of the released information. This is why it has been issued that independence, the first of all ES CoP principles, is at special risk in a situation like this and it has been criticizes that the ever increasing demands are not matched by a sufficient willingness to provide NSIs with (financial) resources (Ljones, 2011).

We conclude that National statistical offices such as Statistics Spain find themselves in the challenging situation of having to meet quality requirements in several dimensions at once and that innovative methods and strategies are required in order to jointly implement those sometimes conflicting quality principles.

2.2 Selective Editing

The totality of tasks performed by Statistics Spain related to the statistical production can be defined in 8 different phases according to the Generic Statistical Business Process Model (GSBPM). The GSBPM is an official standard in the area of statistical production, which describes the necessary activities and tasks to transform raw data into statistical information (UNECE, 2019a). The eight phases of the statistical production process are to specify needs, design, build, collect, process, analyse, disseminate and evaluate. Each phase is composed by different subprocesses. These are arranged in a logical and typical sequence, but they do not represent a linear process, nor do all steps necessarily have to be carried out. The GSBPM rather aims to provide a guideline including all the possibly necessary tasks, which can be recursive.

2.2.1 Data Editing within the Statistical Production Process

Out of the eight phases, the subject of this thesis is related to the fifth phase (*process*). The processing phase is made up of numerous sub-processes and its standardization is of particular difficulty, due to the high degree of heterogeneity which arises from the fact that processing tasks are applied to a huge variety of (survey) data and are carried out by different offices and sections inside Statistics Spain. Data processing describes the cleaning of data observations and their preparation for analysis. To this end, data has to be revised and cleaned, errors must be detected and edited, so

that new variables can be derived from the data, weights can be calibrated and totals can be estimated.

More precisely, selective editing is carried out in the subprocess 5.1 *Edit & Impute* (UNECE, 2019a). As part of data processing, editing is a crucial and resource-consuming part of the work of the national statistical offices. A state-of-art study among statistical agencies carried out in 2007 found out that about half of the respondents spend more than 50% of their overall resources on E&I (Luzi et al., 2007, p. 50). The incoming raw data usually contains erroneous values, which must be detected and treated so that the derived aggregated data, that will later be disseminated, meets the quality standards. Error treatment in the course of the data editing process has been done exclusively manually for many years, and although that is not the case anymore, manual error detection and recontacts/follow-ups are still widely used. In manual editing the correction of erroneous values is archived by recontacting the reporting entity (e.g. companies). However, recontacts increase the response burden and are also associated with slowing down the editing process (De Waal et al., 2011). This is one reason that makes the editing process extremely time and resource consuming (De Waal, 2013). On the one hand, the detection and treatment of errors are essential steps for the accuracy and reliability of the outputs; on the other hand, this time-consuming practice collides with other quality dimensions described above, such as cost effectiveness or timeliness.

Selective Editing are techniques derived from the need of balancing these different quality requirements. Over the years, the statistical offices have learned from their experience that particularly observations can be identified which are responsible for much of the inaccuracy in final estimates. This means that not all errors need to be corrected (Granquist, 1997), however reluctant this may sound at first in the context of producing high quality statistics. The errors that indeed need to be treated are described as the so-called *influential errors*, which are found to have a substantial impact on publication figures for those variables (Luzi et al., 2007). Selective editing techniques rank observations according to their potential to contain influential errors in order to select units for further interactive editing.

We keep in mind that selective editing is one of various E&I (Editing and Imputation) strategies, that can be used in combination with other editing techniques to make the data editing process as efficient and quality-assuring as possible. As visualized in figure 2.1 below, selective editing is applied after an initial E&I, and is followed by interactive E&I applied to the units that are detected as influential, as well as an automatic E&I step that creates the micro-edited files, before macro editing is applied [3]. The selection of influential errors for interactive treatment is a so-called control element in the process flow, as it defines the conditions for deciding which of the alternative following tasks are carried out (UNECE, 2019b).

[3] For an overview of these editing modalities, see for example De Waal et al., 2011.

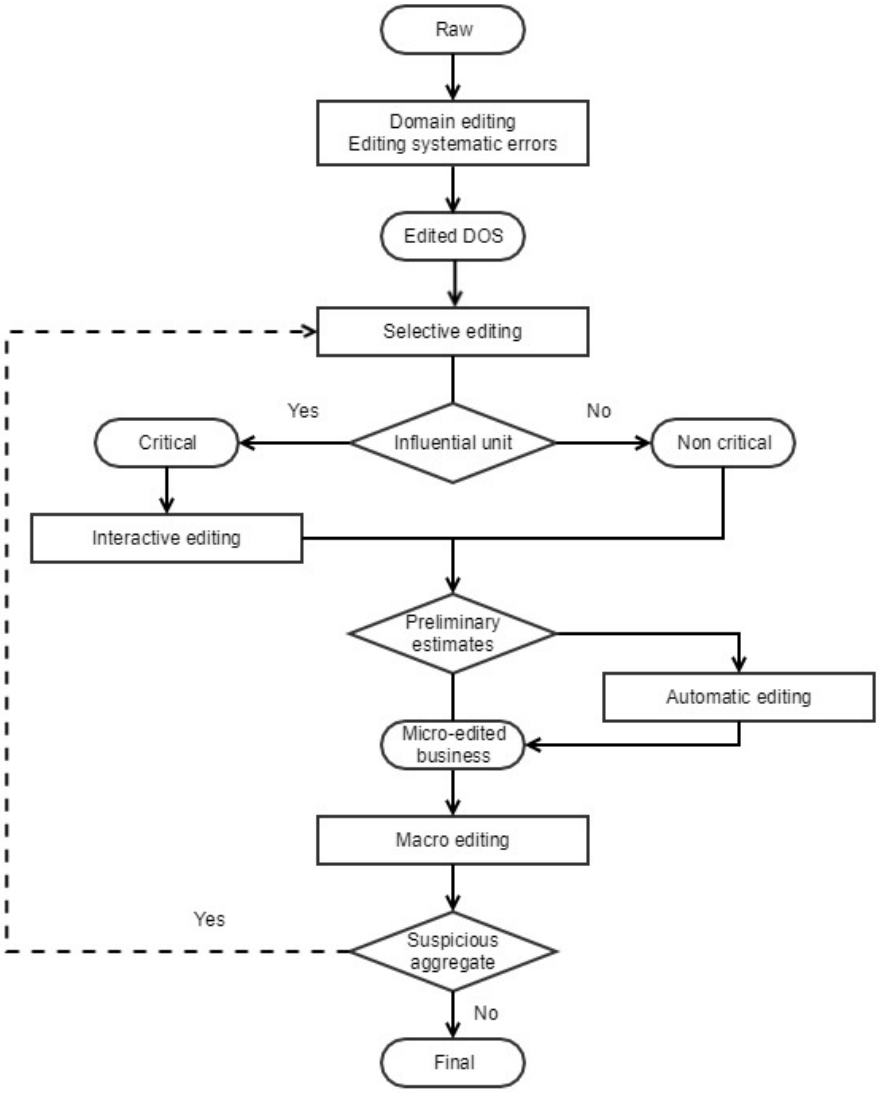

FIGURE 2.1: SDE Flow Model for Short-Term Business Statistics (UNECE, 2019b)

As described by Pannekoek et al. (2013, p. 523) and later outlined in the Generic Statistical Data Editing Model (GSDEM) of the UNECE, the activities related to the data editing process can be described as statistical functions to provide a common terminology. Considering the three statistical data editing functions "review", "selection" and "treatment", that group activities to be carried out in the E&I process, selective editing can be categorized as a review function, and combined with a set threshold as a selection function, because it selects specific fields or units for further treatment (UNECE, 2019b).

2.2.2 Score Functions as Instruments of Selective Editing

So how are influential units actually detected? The score function approach is a way of implementing the basic idea of selective editing: reducing resources allocated to interactive manual editing without compromising accuracy. Score functions are a tool to prioritize units according to the expected importance of inspecting it, that is, the expected benefits of correcting errors in this unit (Hedlin, 2003; Scholtus et al., 2014). To calculate this score for a unit as a whole (the global score value), local score

values are computed for all the relevant items of a unit. The generic way of creating a global score function for a unit k and quantitative variable y can be described in four consecutive steps:

1. Computation of anticipated values of the variables of interest $\hat{y}_k$
2. Computation of local (item) score values s_k
3. Computation of global (unit) score values S_k.
4. Determination of threshold values C_k .

1. Item score functions are based on the relation between the raw value of the variable y and the anticipated value $\hat{y}$. Compared to e.g. imputation models, the computation of anticipated values in this context is usually of rather low quality, it could for example just be defined as the edited last month value of the variable.

2. Based on this value, the local (item) score functions are created. It has the following generic form:

$$s_k(y_k, \hat{y_k}) = F_k(y_k, \hat{y_k}) x R_k(y_k, \hat{y_k}),$$

representing both the risk component R_k and the influence component F_k (De Waal et al., 2011). The former refers to the likelihood of a potential error while the latter refers to the contribution of unit k to the final target estimate. Between the different possibilities to calculate the score functions, some of them explicitly include the risk component as the probability of a value to be erroneous, while other approaches are limited to the comparison of observed and anticipated values. It has been found that the latter tend to produce high rates of false alarms (Scholtus et al., 2014).

3. In a third step the local score functions are combined to a global unit score function. To compute the unit score function different weights can be assigned to the local scores which are associated to different variables that might be considered less or more important for the global score. There are again different functions that can be used to generate the global score values. With s_{kj} being the item scores of item j of unit k, simple ones would be for example the sum of item scores function $S_k = \sum_j s_{kj}$, the max function $S_k = \max_j s_{kj}$or the euclidean score. Those can in turn be unified for example by one of the more sophisticated Minkowski functions of order α , which can also include a weighting w_p (Hedlin, 2008):

$$S_k^{(\alpha)} = (\sum_{p=1}^{P} w_p [s_k^{(p)}]^{\alpha})^{\frac{1}{\alpha}}, \quad for \quad \alpha \geq 1$$

4. Based on the global score, units are ordered according to their expected impact on target estimates. A threshold C_s is established to select those observations with a score value above this threshold for interactive editing. Thresholds can be set based on the desired level of quality, so that errors in not selected units only have a negligible influence on the publication cell aggregates (Di Zio and Guarnera, 2013). Simulations studies, based on unedited raw data and a fully manually reviewed data set, are usually carried out to find an appropriate threshold value. To this end, after calculating the global scores for the raw data, it is simulated that only the first n observations are manually reviewed by replacing their raw values by the edited

data, while the rest of the records keeps their raw values. This procedure is repeated for different values of n. The target measures are then calculated for the mixed data sets with n edited units and compared to the target measure of the fully edited data (De Waal et al., 2011).

At Statistics Spain, in particular from 2013 onwards, efforts have been made to revise the editing and imputation strategies according to the frameworks proposed in the EDIMBUS manual as part of a modernization process. Efforts were made to parameterize the editing strategies and to extend the EDIMBUS structure, for example by including editing tasks performed already during data collection (Rama and Salgado, 2014). In this context, Arbues et al. (2013) have developed an approach that considers the minimization of editing resources while assuring data quality in the selective editing process from the perspective of a mathematical optimization problem. An optimization problem is the problem of finding the best solution from all feasible solutions and seeks to minimize a loss function. This functions maps values onto a number intuitively representing some "cost", in this case, manual resources and loss of quality. In the approach, different kinds of optimization problems are formulated, depending on which kind of auxiliary information out of the three types longitudinal, cross-sectional and multivariate is used. According to this approach, the local score functions are calculated as a conditional expectation in terms of the auxiliary information available. The optimization approach to selective editing has been implemented for example in the E&I strategies of the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey (López-Ureña et al., 2014).

The idea followed in this thesis joins the quest for an efficient data editing phase and proposes the application of random forest algorithms in the construction of score functions. A score functions can be naturally understood as "an estimate of the error affecting data" (Di Zio and Guarnera, 2013). Random forests can be used to model these errors between raw and anticipated values, on which the local score functions are based. The approaches used in this thesis will be explained in the following section, before the SSAI survey data are introduced as our data basis in chapter 3.

2.3 Approaches to the Application of Random Forests in Selective Editing

The two approaches described in the following are inspired by the optimization approach to selective editing (Arbues et al., 2013), mentioned in the previous section. Since the error between raw and edited turnover values

$$e_k = y_k^{raw} - y_k^{ed}$$

is a numerical quantity, the initial purpose of this thesis project was to use regression forests, i.e. random forests with a quantitative target variable. To train the forest, the available data was divided in a training set and a test set based on the corresponding months the data originates from. The local score function was expressed as a conditional expectation in terms of the auxiliary information $\mathbf{Z}_k^{\text{cross}}$ available:

$$s_k = d_k \cdot \mathbb{E}\left[|y_k^{raw} - y_k^{ed}| \big| \mathbf{Z}_k^{\text{cross}} \right], \tag{2.1}$$

where d_k is the design weight of the unit k.

The idea was to predict the error in turnover of the test month based on the raw turnover values and the auxiliary information. However, the first attempt to define an error variable and try to adjust a random forest didn't lead to the expected results, and prediction accuracy fell short of expectations. For a detailed analysis of the steps and results of this first approach, see chapter 5.

After a closer analysis of the characteristics of the data and the distribution of the error a second, more sophisticated approach was developed. The first methodological decision was to treat separately the observations with missing values in the turnover variable, on which the target variable is based. For the rest of the observations we implement a two-step random forest approach to accommodate the semi-continuous distribution of the variable, which turns out to have a large number of observations with value 0 and a small continuous part.

We define $e_k^{\text{obs}} = \delta_k \cdot \epsilon_k^{\text{obs}}$, where δ_k is a Bernoulli variable with values `1`, if the observation is erroneous (with a probability p_k) and `0`, if the raw value is correct, e.g. the error is 0 (with probability p_k). The first step uses a two-class decision forest to determine if an observation is predicted to be class 0 or class 1. For the second step, the continuous variable ϵ_k^{obs} represents the magnitude of the error in case it is produced. To estimate this second step target variable, a regression forest is used, given that the output variable is continuous. The idea is to use *random forests* in two steps to model these two random variables based on the available raw information as follows:

1. A binary variable representing the unit type is constructed to model δ_k by classifying units as type *correct* or *error*. Then, a random forest can be built to model this variable and the class probabilities p_k are computed using the available auxiliary information.

2. To model $|\epsilon_k^{\text{obs}}|$ we build the second step random forest based on the subset of the data that is of type *error*.

It had to be kept in mind that this approach substantially reduces the number of observations for the regression forest, which is why the data basis is enlarged by data from previous months and a longitudinal information is added to the auxiliary information Z_k^{aux}. Besides, the target variable of the classification forest is highly imbalanced. Solutions to this problem and the related methodological decisions are described in chapter 5. The two models are now described by $\hat{p}_k \equiv \mathbb{E}[p_k|Z_k^{aux}]$ and $\mathbb{E}[|\epsilon_k^{\text{obs}}|\big|\delta_k = 1, Z_k^{aux}]$. Hence, the constructed score function s_k is expressed by

$$\begin{aligned} s_k &= d_k \cdot \mathbb{E}\left[\delta_k^{obs} \cdot |\epsilon_k^{obs}|\big|Z_k\right] \\ &= d_k \cdot \mathbb{E}_{\delta_k}\left[\mathbb{E}_{\epsilon_k}\left[\delta_k \cdot |\epsilon_k^{obs}|\big|\delta_k^{obs}, Z_k^{aux}\right]\right] \\ &= d_k \cdot \hat{p}_k \cdot \mathbb{E}_{\epsilon_k}\left[|\epsilon_k^{obs}|\big|\delta_k^{obs} = 1, Z_k^{aux}\right], \end{aligned} \tag{2.2}$$

with $\hat{p}_k$ being the estimated probability that observation k has an erroneous turnover value (unit type *error*).

Chapter 3

Data basis: Short-term Business Survey Data

The analysis related to our selective editing approach will be carried out based on data from the Services Sector Activity Indicators (SSAI). The SSAI survey is a short-term business statistic conducted by Statistics Spain (2020).

Short-term business statistics are aimed at providing information on the business cycle of an economy and are usually carried out monthly or quarterly. In contrast to household surveys, business surveys are characterized by a relatively small number of variables which use to be mainly numerical. Units can be difficult to delimit due to their complexity and evolving nature. The population size is rather medium or small for business surveys (Scholtus et al., 2014). Due to the short term cycle, time series data from previous waves of surveys are quickly available for short term surveys, even if the survey has not been implemented for a long time. For short-time business surveys, results are disseminated relatively shortly after the data collection, in case of SSAI 51 days after the last day of the reference period on average (Statistics Spain, 2019). Disseminating with the shortest possible delay is necessary to ensure the relevance and timeliness of the published data, but also sets a limit on the time that can be invested in data editing, making an effective data editing process essential. De Waal et al. (2011, p. 6) state that due to numerous edit rules and relatively large numbers of errors compared to social surveys, business surveys are associated with a especially high effort in data cleansing. On the other hand, they also have a characteristic that makes the use of selective editing particularly fruitful, namely the usually large skewness of distributions of the important variables. That means that a few observations have a major impact on the aggregated results (such as companies with high turnover and large number of staff), while the impact of smaller companies is comparatively low. This benefits the implementation of the fundamental idea of identifying influential units and makes selective editing an especially attractive technique.

3.1 Data basis: The SSAI Short-term Business Survey

The Services Sector Activity Indicators SSAI (IASS in spanish, from 'Indicadores de actividad del sector servicios') measure the short-term evolution of the activity of companies operating in the non-financial market services in Spain. The activity indicators, which are reported in nominal terms, are based on two main variables: turnover and employed personnel. The variable *Turnover* consists of the amounts invoiced by the company, during the reference period, for the provision of services and the sale of goods. *Employed personnel* comprises both wage-earning and unpaid personnel like working family members (Statistics Spain, 2019, 2020).

In order to obtain this data, an ongoing survey is conducted, collecting data from more than 28,000 companies that operate in the sector every month. SSAI provides information as to CNAE 2009 classification of economic activities[1]. The population scope is made up of companies whose economic main activity is classified as Trade, Transport and Storage, Accommodation, Information and Communications, Professional, Scientific and Technical Activities or Administrative and Support Services Activities (sections G to J, N or M in the CNAE 2009 classification). Units are selected out of the frame via stratified random sampling (Statistics Spain, 2019, 2020). The survey was first launched in 2000 and is available for all divisions since 2002. Since then, some methodological changes have been made, such as the expansion of the sample size in 2005 in order to be able to publish results at regional level. Since 2013, results are presented as chain-linked Laspeyres indices, in order to measure variations as compared to the base year. As of the reference month January 2018, indices are calculated and published in base 2015, in order to comply with the requirements established in Eurostat Regulation 1165/98 on short-term statistics (Statistics Spain, 2019).

3.2 Data Collection Process

In the case of the SSAI data collection methods of primary (direct) data collection are used, as it is a survey. The data are collected by completion of the questionnaire by the respondent, using one of the following methods: Internet (via the online tool of IRIA system), e-mail, fax, telephone or postal mail, so we are looking at a mixed form of data collection (Statistics Spain, 2020). As we can see, the actual SSAI survey consists of only a few variables, basically the turnover, the number of unpaid personnel and the number of paid personnel, differentiating between employees with fixed or temporary contracts. Nevertheless, during the collecting and editing process, a lot more (meta-)data are generated and stored in internal variables, which will be called paradata variables. Like described in chapter 2, the data which is collected and recorded by the regional offices, will be already edited in the collection process. The paradata therefore contains information like the date when data was recorded, when it was edited, if specific comments where added in the editing process or by which data editing actors the data was reviewed. The information from these survey variables is furthermore complemented by identification information, for example about the branch according to the CNAE 2009 and the autonomous community which the units belong to, etc.

Statistics Spain stores variables from different statistical operations and related information in a data repository which is a key–value store, in order to have a standardized database providing the necessary data for the data editing phase or any other internal operation based on those data. The repository contains files with different kinds of information related to each statistical operation. More precisely, it contains files with the recorded unedited raw values (FG), files with the values, that have been edited in the field but not validated by the data managers (FD) and files with the final microdata after data editing has been performed (FF). This final version of the data is the basis for the aggregates and indices disseminated by Statistics Spain. In addition, there are files that contain the value of paradata variables (FP), the value of a direct identification variable (FI), the value of a validation interval for editing during collection (FL) and a value for the cross-sectional selection of units with influencing errors (FT).

[1] National Classification of Economic Activities, Spanish version of the NACE Rev.2

Variable name	**Description**
	Survey variables
b1_raw_t	Raw turnover value
c11	Unpaid staff
c121	Paid staff with fixed contract
c122	Paid staff with temporary conctract
b1_raw_t_1	Raw turnover value from last month
b1_ed_t_1	Edited turnover value from last month
	Identification and additional variables
existencias	Value of inventory
exist	Dichotomous variable, presence (1) or absence (0) of existencias
cnae	CNAE 2009 Classification
CCAA	Autonomous Community
CodProvincia	Identification code for the province
CodTame	Code for tame
rama	Internal economic activity classification similar to CNAE
factor	Survey weight of the unit
	Paradata variables
codUGestion	Code for the unit responsible for data collection
varGestion	Variable related to the data administration
CodAgente_1	Code for the agent responsible for recording each unit
Usuario	Code for the agent that added comments
fechaRecepcion	Date of data reception
fechaGrabacion	Date of recording data
fechaDepuracion	Date of editing data
observaciones_X	List of variables containing various types of comments related to the E&I process, like recontacts, observation of irregularities, etc.
	Derived variables
dias_Rec_Grab	Number of days between data reception and recording
dias_Grab_Dep	Number of days between data recording and editing
early_Rec	Dichotomous variable indicating if data was receipt in the first half of the month (1) or not (0)
early_Grab	Dichotomous variable indicating if data was recorded in the first half of the month (1) or not (0)
early_Dep	Dichotomous variable indicating if data was edited in the first half of the month (1) or not (0)
CodProvincia1	First digit of identification code for the province
CodTame1	First digit of identification code for the tame
cnae1, cnae12, cnae123	First, first two, and first three digits of identification cnae classification code
b1_relDiff_1	Relative change in turnover compared to the previous month
errorb1_abs	Absolute error in turnover between raw and edited turnover value
b1_error	Dichotomous variable indicating if the raw turnover value differs from the edited value (1) or not (0)

TABLE 3.2: Variables used in the random forest models

In the context of this thesis, four different types of variables will be used for the selective editing algorithm: (1) survey variables, (2) identification variables, (3) paradata variables from the editing process and (4) derived variables which are generated based on the previous variables in order to increase the performance of the selective editing algorithm. The variables used in the selective editing analysis are listed in Table 3.2. In addition, some longitudinal variables are added, even though most of the variables are from the reference month. To compute the continuous target variable, the edited turnover value from the final files is used to calculate the error in the turnover value.

$$|errorb1_t| = |b1_t^{ed} - b1_t^{raw}|$$

. The second target variable is created by dichotomizing the former, based on the presence or absence of an error.

$$b1_{error} = \begin{cases} 0 & \text{if } |errorb1_t| <= 0.01, \\ 1 & \text{if } |errorb1_t| > 0.01 \end{cases}$$

In this thesis, SSAI data from the September 2019 until March 2020 are used. The random forest built in the first approach is based on the data of only one month, namely the SSAI data from February 2020, while the data from March is reserved as test data. In the second approach, due to the reduced sample sizes resulting from the two-step procedure, the data basis was expanded with information from further months and longitudinal information was included. Therefore, data from month of September was only used to derive the longitudinal $b1\_relDiff\_1$ variable, the months October 2019 until February 2020 served as training data, and March was again used as test data.

3.3 Data Maturity

In a short discussion we would like to address the issue of data maturity. While we have seen above that there are strict requirements and standards regarding interpretability and standardisation for published information and variables, internal variables, which we use in the random forest models, are generated based on the needs and purposes of each department or unit. Although the data repository attempts to standardise the information and paradata it contains, many of the variables and their background are not self-explanatory and they result difficult to interpret and use for staff that is not familiar with the details of the steps carried out in each operation.

Two major problems seem to be 1. the history of the paradata (it is not clear when and for what purpose information was created) and 2. the documentation of the data, codebooks are still incomplete and for some internal variables no meta-information is available. In addition, new types of variables are added over time, while others disappear, but this fluctuation is not explicitly documented. This is particularly problematic for the application of automated procedures, as the advantage of self-improvement is lost if these circumstances mean that the data basis must repeatedly specified manually.

Chapter 4

Machine Learning and Random Forests in Official Statistics

In this thesis, random forest algorithms were selected from a variety of possible machine learning methods to address the presented research question. This chapter contains introductory remarks on what is machine learning, a short comment on the application of machine learning in official statistics, it discusses the most relevant methodological aspects of random forests, and presents two different approaches of applying these for selective editing in the context of the data used in this thesis.

4.1 Short notes on Machine Learning

Machine learning (ML) generally describes "a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" (Murphy, 2012). So the main idea is to use algorithms to parse the data, learn from it, and be able to make suggestions or predictions about something. The flexible approach of ML makes it fit to analyse large amounts of data, to deal with uncertainty and to process high dimensional data (Barber, 2012). Since data processing technology is providing increasingly better resources for machine learning, e.g. through extended computing power, the popularity and application of these techniques have increased enormously in recent years. Its advance is not limited to a specific field or discipline, but has a huge range of applications.

Machine learning techniques can be divided into two main types, namely supervised and unsupervised learning techniques. In supervised learning, each observation is associated with a response y_k, the target variable. The aim is to model the relationship between the response and the measurements of the predictors, or predict the response for new observations (James et al., 2013). One is therefore interested in methods that work accurately with previously unseen data (Barber, 2012). Unsupervised learning is even more widely applicable, as there is no need define a desired output. Unsupervised learning algorithms search for patterns based on the set of features of interest in the data, but without a labelled response variable. This style of learning, of discovering knowledge, is sometimes compared to the way that humans and animals learn. To reduce complexity and to make it easier, techniques like PCA can be used which reduce the dimensionality of the data (Murphy, 2012). From a probabilistic point of view, supervised learning can be understood as conditional density estimation, while unsupervised learning is described as unconditional density estimation. Supervised learning usually only tries to predict one variable, which is why it corresponds to univariate probability models, while unsupervised learning would require multivariate probability models (Murphy, 2012).

A common problem mentioned in the context of machine learning is *overfitting*. An overfit model is one that is based too strongly on the observed data and therefore performs well only on the training data but can't produce accurate results for previously unseen data. Model performance should therefore also always be assessed when applied to unknown data. To gain insights about the quality of the models already in the training phase, before choosing a final model which is applied to test data, resampling methods like bootstrapping or k-fold cross validation can be used. Resampling models repeatedly fitting a model while intentionally leaving out part of the data, which can then be used to analyse the model performance (Boehmke and Greenwell, 2019).

4.2 Machine Learning in Official Statistics

In recent years the application of Machine Learning techniques has also become more and more popular among statistical agencies. Although the different statistical institutions face a variety of difficulties and business environments, they share common types of problems. In the last years ML methods started to be widely recognized in the context of the official statistical production and statistical agencies increasingly seek the advantages of their applications, like their high potential to increase efficiency, reduce response burdens and emerging opportunities to use new digital data. The UNECE High-Level Group for the Modernisation of Official Statistics has included Machine Learning as one of its modernization projects in 2019, which is being continued in 2020. The potential use of Machine Learning for Official Statistics was recently addressed in a webinar series of the UN World Data Forum in July 2020 entitled "Adding Value to Statistical Data Production through Machine Learning". Within this framework, several Machine Learning sprint events have been organized and progress reports have been published in the last two years. In the context of these initiatives, various pilot projects were carried out and experiences of different statistical authorities were collected mainly in three areas: Classification and Coding, Editing and Imputation, and Imagery (Julien, 2019).

As an example, the US Bureau of Labor Statistics uses Automated Coding for their Survey of Occupational Injuries and Illnesses in which text entries have to be converted into standard codes. These tasks, which had been done completely manually until 2014, could be automatized thanks to a machine learning model, which was trained with a subset of features, such as words or pair of words and their related codes. Not only did both developed autocoders, one based on logistic regression and another on based on a deep neural network, save a massive amount of resources (manual effort, time). Contrary to the reservations that machine learning algorithms could worsen the quality, it could also be shown that the autocoders based on ML methods are more accurate than manual coding (Measure, 2017). Criticism related to machine learning also states that its methods have a black box character and are often uninterpretable (Molnar, 2020). This aspect must always be taken into account in the context of official statistics in order to maintain the trust of its users. Another bottleneck that seems to slow down the application of ML in official statistics is a lack of expertise and experience. Even if many ML methods and advantages are superficially known, those responsible may lack the time to familiarize themselves sufficiently with the subject area and find areas of application in their own area of responsibility (Beck et al., 2018).

In a 2018 study, staff members of the Federal Statistical Office of Germany investigated to what extent the application of machine learning is widespread in their

own statistical institutions, in the national statistical offices of the European countries, as well as in other selected countries such as Canada or the USA (Beck et al., 2018). They have also looked into the question of which techniques are used and for which areas and tasks they are applied. For this purpose, 39 international statistical institutions were contacted, the majority of which are NSIs from the EU Member States. It is found that a considerable proportion (21 of 33) of the institutions that provided an answer, is already involved in machine learning projects, although they find themselves in different project phases. Most of the reported 136 projects (45%) are still in an experimental stage, while about 19% are in development and 15% are already in productive use. The remaining 21% of the projects were only formulated as ideas in that moment (Beck et al., 2018). As for the machine learning methods that were applied, random forests were by far the most common method, but neural networks and support vector machines and other decision tree methods were also found to be very popular. The implementation of ML techniques can be found across a variety of tasks among various phases of the GSBPM (Design, Collect, Process, Analyse, Evaluate), but especially in the process phase. The most frequently mentioned application types were automatized classification, the imputation of values and microdata linking.

Despite the widespread use of ML techniques in E&I strategies, most of the examples are related to the imputation of values and relatively little can be found about the use of ML for the detection of non-sampling errors. A few projects of this kind can still be found: The Federal Statistical Office of Switzerland was testing different ML methods such as generalized boosted models, Random Forest, Neural networks, Naive Bayes or Tree algorithms to perform the detection of suspicious responses. Observations are classified as suspicious or non-suspicious units which the aim of recontacting those where anomalies were found. At Statistics Norway, Classification by Random Forests are explored for editing purposes in register based salary statistics (Beck et al., 2018). Some more generalized software solutions that also include the detection of suspicious units have been developed. One example is CANCEIS, which was developed by statistics Canada, but is already experimented to be used also in other statistical agencies, like for example Stats NZ or the Federal Statistical Office of Germany (Lange, 2020; Spies and Lange, 2018; Stats NZ, 2019).

4.3 Random Forest Methodology

Random forests are a non-parametric machine learning method first introduced by Breimann (2001). They are a supervised learning method, which means that both input and output are explicitly indicated in the training data. The learning set D is therefore a set of pairs of input vectors and output values $D = \{(\boldsymbol{x_1}, y_1), ..., (\boldsymbol{x_n}, y_n)\}$ where $\boldsymbol{x_k} \in X$ and $y_k \in Y$, with X being the input space, Y being the output space and N being the size of the training set.

4.3.1 Construction of Random Forests

Random forests are based on decision trees, which can be either regression or classification trees. Therefore, they are an ensemble method, which combines information of multiple models, in this case within a specific technique, decision trees.

Decision trees are a tool to predict the generally consist of a root node, which is the top node, any number of internal nodes and at least two leaf nodes, which are the terminal nodes. Nodes can be understood as questions that are asked and which

classify an observation according to its answer. These can be categorical questions like the CNAE of a company, or questions related to a continuous variable, e.g. if a company has more than 100 employees or not. Based on the answers, each observation will end up in one of the leaf nodes. Each leaf node is associated with a value of the target variable. This value is derived from the values of all the observations of the training data which belong to each leaf node, usually the mode of the categories in case of classification, or the mean value of a numeric values in case of regression trees. This value will therefore be the predicted value for a new observation that ended up in the corresponding leaf node (Moisen, 2008). Decision trees are built by recursive binary splitting or partitioning, which means that the feature space is split up by splitting a node into its two descendants. Which split is chosen depends on the evaluation of all possible splits of all the predictor variables which is based on a criterion like the residual sum of squares for regression or a node impurity measure like gini index for classification (Cutler et al., 2011; James et al., 2013).

Classification and regression trees (CART) have a number of attractive advantages such as, as their name suggests, that they can be used for the prediction of both categorical and numerical variables. Predictors as well can be categorical or numerical variables or even a mixture of both. CARTs are said to be easy to build, easy to use, and, if the trees are small, they are intuitively interpretable, a property that is hard to find among machine learning algorithms. Moreover, they inherently perform a feature selection, whereby irrelevant variables are basically filtered out (Hastie et al., 2009, p. 352). However, decision trees have a weakness that prevents them from being perfectly suitable and from being able to compete with other machine learning algorithms, namely accuracy. Their predictions are related to high variance, which makes them insufficiently robust.

Random Forests solve these problems in two ways. It is known that high variance low bias problems can be solved by bootstrap aggregating (*bagging*). If this technique is applied to decision trees, it means that not only one tree but a whole forest is constructed. The target variable to be estimated is then obtained by aggregating the results of all trees. To be able to build several trees from the same sample, subsamples are drawn with replacement during the bagging process, so that each tree is based on a slightly modified data basis. This results in more divers trees and the variance is reduced on average. Nevertheless, the trees will still be very similar, as they all take the same set of variables into account. This will be especially problematic when a few predominantly important variables are present in the data, which are then virtually always selected.

It has been shown that the generalization error in case of random forests is based on the quality of the individual trees and the correlation between the trees in the forest (Breiman, 2001). The generalisation error for a supervised learning model m such as a random forest algorithm can be expressed by

$$Err(m_D) = \mathbb{E}_{X,Y}\{L(Y, m_D(X))\},$$

with L being a loss function based on e.g. the mean squared error for regression or another measure of discrepancy between its two arguments. In order to prevent the correlation between the trees and decrease the error, random forests use a second technique, whereby they do justice to the "random" in their name: For the evaluation of each split, a new random sample of m predictor variables is taken into account as eligible variables [1]. In this way, not all trees will be able to make a similar selection

[1] Other ideas of introducing randomness, like for example to build each whole tree on a random selection of predictors, have also been discussed (Ho, 1995).

and will be decorrelated. The two basic approaches of bagging and introducing randomness in feature selection make random forests more robust and stable. On the downside, random forests are less interpretable than simple decision trees. Still, their flexibility and simple application makes them one of the most popular machine learning models.

4.3.2 Features and Properties

The characteristics of the random forest will now be explained a little further before we address hyperparameter tuning.

Out-of-bag Validation. Validation is an important aspect when constructing random forest models. It is not a good idea to base the evaluation on training data itself, as it can lead to selecting a model that overfits the data. some of the available data has been put aside as test data, but these are first used to verify a pre-selected model. Other mechanisms should therefore be used during the modelling process. Resampling methods like k-fold cross validation or bootstrap can be used for this. Both don't require leaving out an extra part of the data for validation (Boehmke and Greenwell, 2019). An advantage of models like random forests that build bootstrap samples, is that the out-of-bag (OOB) samples, e.g. those who were left out at the construction of a particular forest, can be used to estimate the generalization error by using trees for which the observation is out-of-bag for the prediction of the response of a units (Cutler et al., 2011). Of course, other validation methods like cross-validation or bootstrap can also be used. Concerning the question which of the three methods is most recommendable, it has to be said that the computational cost of OOB validations is drastically lower, since these OOB samples, as mentioned, already exist as in-built feature of the method. Kuhn (2013) find that there are very little differences between using OOB or 10-fold cross validation in a study they did based on CART trees. Comparing cross-validation and bootstrapping, another study finds that K-fold CV tends to have higher variability than bootstrapping. Because of repeated observation, bootstrapping can increase the bias of the error estimate. However, this issue is usually negligible with large data sets (Boehmke and Greenwell, 2019).

Variable Selection and Importance. The selection of suitable variables is always a difficult task in the modelling process, especially in new areas and when one can only rely on theoretical assumptions to a limited extent. Decision trees, as mentioned above, have the excellent property of automatically selecting variables in the modelling process and are thus reasonably resilient to the inclusion of irrelevant predictors. When working with a collection of bagged trees in random forests it becomes impossible to understand the relation between the response and the predictors (James et al., 2013; Kuhn and Johnson, 2013). However, we can still evaluate the importance of each predictor by summarizing information from the trees in the forest: For regression trees, we can average the total value that the RSS decreases due to splitting with a predictor p_k over all B trees. Equivalently the reduction of the Gini index can be averaged over all trees to determine the importance of a variable in a classification forest. It should be kept in mind that strongly correlated variables can lead to the overestimation of the importance of a variable that is not important but correlated with an important variable. The number of split variables also has an important influence on the importance values (James et al., 2013). Even if the forest deals flexibly with irrelevant information, it can be useful to filter out so-called

near-zero-variables from the model in the preprocessing process. Some packages offer this option as an automatic option before the model building process as it is in easy way of making the model less costly and more interpretable (Boehmke and Greenwell, 2019).

Other characteristics. Apart from being applicable to both regression and classification problems, random forests also naturally handle multi-class problems. They can be used for high-dimensional data and they are scalable to large learning set. Random forest also naturally include a proximity measure, which is based on the number of trees in which two observations end up in the same leaf node compared to all the trees of the forests. Proximity measures can for example be used to detect outliers or to impute missing values. Random forests are furthermore easy to run in parallel, as all trees are created independently from each other (Cutler et al., 2011; Kuhn and Johnson, 2013; Louppe, 2014).

4.3.3 Hyperparameter Tuning

Apart from the characteristic that Random Forests imposes few requirements on the data, the method is at times described as a 'off-the-shelf' tool, insinuating that it can be applied with comparatively little tuning of the available parameters. While Kuhn (2013) and others find that tuning parameters have relatively little influence on the error metric, there are still some parameters that can have an important impact on the outcome and performance of the random forest. In the following we will mention those parameters and discuss how optimal values can be found based on the remarks of Probst (2019).

Number of trees: One important parameter the researcher has to decide on is the size of the forest. By definition, prediction gets more accurate with an increasing number of trees. The number is not a tuning parameter in the true sense, since the mathematical optimum would be an infinitely large number. Nevertheless, growing an unnecessarily big number is computationally expensive, while at a certain point an enlargement of the forest does not further profitably reduce the error, which would be inefficient. It can therefore be reasonable to establish threshold

$$C > \frac{Err_{n_{tree}}(m) - Err_{n_{tree}+1}(m)}{Err_{n_{tree}+1}(m)},$$

where *ntree* is the number of trees in a forest and the threshold C expresses the minimum relative error reduction that we consider worth adding an additional tree. Visual evaluation of the error curve as a function of the number of trees also uses to be a helpful tool to detect when the curve flattens out noticeably. On the other hand, the number of trees to be grown can be limited by external factors like time resources or computing power.

Number of variables to be considered for each node (m_{try}): Like explained before, random forests prevent its trees to be highly correlated by considering only a random selection of variables at each node. The optimal number of variables to be considered has to be found between 1 and the total number of independent variables p available in the data set. If the behaviour of the error was previously analysed it can also be useful to find $min(Err_{m_{try}})$ for $m_{try} \in \{a, b\}$ by establishing a range $\{a, b\}$ with $a > 1$ and $b < p$ to make a hyperparameter grid search more efficient. Taking all variables into account would be equivalent to bootstrap aggregating *(bagging)*. In

the literature they can be found different recommendations about the approximate optimal value regarding the number of variables (p). While some authors recommend $m_{try} = \frac{p}{3}$ other say that the optimal value is likely to be around $\sqrt{p}$, the latest standard is to use the former for regression and the latter for classification forests. Some R packages like *randomForest* have an inbuilt tool to find the optimal value for m_{try} in the concrete data set. This tool starts to calculate the error for an default value of variables and allows to set the relative improvement in OOB error as a threshold for the search to continue.

Minimal node size: To not grow the random forests to its maximum complexity, a minimal value for the node size is usually established, that is, how many observations of the training data have to end up in a leaf node. If the number would be below this threshold, a node cannot be split any further. To reduce the minimum node size reduces the complexity of the tree and lowers computational costs, but specific pattern may not be represented. Increasing the minimum value allows for higher complexity of the trees but carries the risk of overfitting. While under-complex trees tend to have higher bias, over-complex ones should come with a high variance.

Additional parameters can be the splitting rule and the sample scheme: Like explained in the construction of random forest, usually a bootstrap sample is drawn with replacement. However, it is possible to sample a number of observations without replacement. In this case a sample fraction should be established. The default value for the sample fraction when is replacement is disabled is 0.632 in some packages, which corresponds to the percentage of the data that is represented on average in a bootstrapped sample. The splitting rule makes different variations of the RF rather than being a hyperparameter. For classification forests, the minimal gini impurity is often used as a criterion to select a variable for splitting the node out of all variables available. For regression trees, the weighted variance is often used as a splitting rule although it is biased towards variables with many categories, alternatives are p-values from a global test, or randomizing the subset of possible splitting values of each variable with the *extratrees* option, which makes computation more efficient.

4.3.4 Treatment of missing values

When building a forest, possible missing values in the data must be taken into account. When a categorical variable has missing values in the training data, these can just be defined as a new class of the variable. When data is missing in a numerical variable, the forest can't handle it, hence the values must be imputed. Apart from imputing missing values with mean or mode of the respective variable, random forest offer another possibility using the proximity measure mentioned above: A first forest is built based on low-quality mean values. Starting from this, the proximity measures are calculated and the imputed value is replaced by a better estimation based on a proximity-weighted average. This procedure is repeated during a few iteration (Cutler et al., 2011). However, this procedure has not yet been used in this study, it was decided to impute values manually. Missing values of categorical predictors were therefore just assigned to class "*". For numerical predictors, a missing *raw turnover* value was imputed with 0. This imputation is justified by the fact that the edited value is the one that an expert finally assigns. The resulting target variable *error* will therefore be equal to the value of the *edited turnover* value for these observations. Units with a missing value in the edited turnover value are discarded.

Missings were also found in the variables *dias_Rec_Grab* and *dias_Grab_Dep*, which where imputed by 60, which was the maximum value of days observed in the data. Missing values in the variables *c11*, *c121* and *c122* where replaced by 0, a missing sampling weight by 1.

4.4 Computational considerations

The programming tasks related to this thesis were carried out in the programming language `R`. Packages like `caret` or `MLR` offer a set of functions to provide a standardized interface for building machine learning models. They include standard task like regression and classification along with their corresponding preprocessing tasks, evaluation and optimization methods to automate standard tasks. As for Random Forests, the R environment offers different packages, of which one of the most common ones is `randomForest` (Liaw and Wiener, 2002), but there is also a number of other packages available, for an analysis of their advantages and disadvantages see Wright and Ziegler (2015). A problem of some of the older R implementations of RFs is that they were optimized for large samples, but not for working with a large number of predictors and can cause very long computation times. We also experienced this problem, building a forest with 500 trees for all of the approximately 30,000 observations in one month and including about 200 variables has resulted in the computations regularly exceeding 72 hours with `randomForest`. The `ranger` package, which was finally used in this thesis, was developed to overcome this problem and has shown that it is on a par with `randomForest` in terms of out-of-bag prediction error and variable importance results (Wright and Ziegler, 2015). `ranger` has shown to be considerably faster than other packages, especially for dichotomous features, while also using less memory that e.g. `randomForest` (Wright and Ziegler, 2015). Similar calculations (with about 30,000 observations) could now be finished in a few hours. However, these problems are mainly a challenge in the design phase, when many forests have to be grown in order to evaluate tuning parameters and are less impactful once the final models have been selected.

Chapter 5

Results of Random Forests applied to Selective Editing

As mentioned in chapter 2, two approaches have emerged from the original problem definition. Therefore, the proceeding and results of the initial approach are presented first. In the course of this, emerging problems will be discussed and the methodological decisions that finally led to the development of the second approach will be explained.

5.1 Simple regression forest

In this first part of the analysis, a random forest model was applied, with the absolute error of the turnover *errorb1_abs* as the target variable. As the target variable is continuous, we model a regression forest. In the course of the search for a suitable model, different variants were analysed, differing as well in the variables included in the model and in the choice of parameters.

In a first variation of the forest, dichotomized comment variables for each individual user were included in the model, indicating if a particular user made a particular type of comment or not. Since in this case almost all of these variables became near zero variance variables and did not turn out to have any importance, we decided to aggregate the number of comments of each type. The dichotomized comment variables are referring only to a certain unit and show if a comment was made without containing information about which user was involved. At this point, no longitudinal information was included in the model.

5.1.1 Random Forest based on Cross-Sectional Information

To get a first impression of how many trees are necessary, quality measure curves were computed based on default regression forests models, analysing the goodness of the model by computing the performance measures by the number of trees. The used measures are the root mean square error (RMSE), the mean absolute error (MAE) and the squared correlation between the observed and predicted values R^2. As trees in a forest are grown one by one, these measures can be easily computed. Given that the error naturally decreases, the objective of the generated graphs is just to give a visual hint on how many built trees are worth the computing power and time.

The graphs in figure 5.1 show the evolution of the RMSE and the MAE as the number of trees is incremented[1]. As it appears, the reduction of the error due to an increasing number of trees begins to stagnate from about 400 trees onwards in

[1]For related additional graphics, like the R^2 curve, please see Appendix A

both measures, even a little earlier in terms of the RMSE. We will therefore build regression forests with 400 trees in order to keep the computational effort as low as possible while still being able to expect decent results.

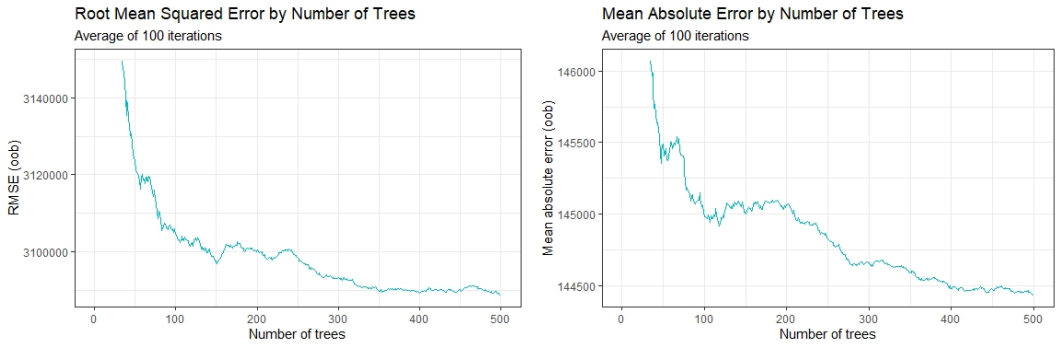

FIGURE 5.1: RMSE and MAE by number of split variables (m_{try})

In a similar manner, we are analysing the quality measures for different values of m_{try}. While analysing the number of trees was aimed at finding a minimal number of necessary trees to reach good performance, the goal of analysing m_{try} is to be able to limit a range of well-performing values, so that the values can be selected more specifically in the search grid. A grid contains all possible combinations of values defined for the tuning parameters represented in it with the aim of comparing the model performance of the different models which result from the grid combinations. Hence, to limit the value range of m_{try} was especially useful to reduce computation time when a high number of variables was used.

Looking at figure 5.2 regarding the number of split variables, it seems like in this case lower errors can be expected for smaller values m_{try}. We will therefore concentrate on a range of low m_{try} values and space them out evenly between 1 and 20 in the grid, which will be built to compare models with a variety of parameter combinations.

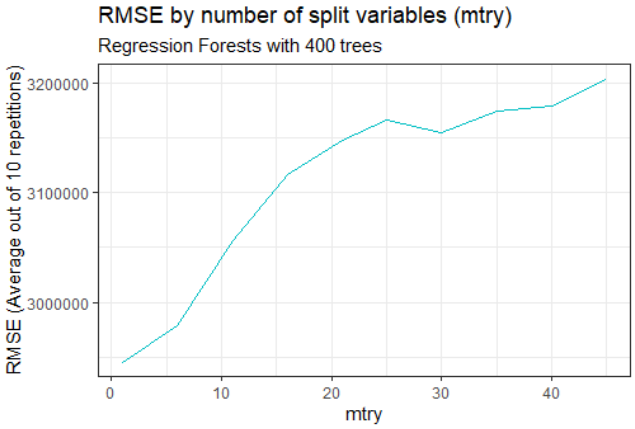

FIGURE 5.2: RMSE by number of split variables (m_{try})

The first forest building process is based on a sample of size n = 27585 and 45 predictors. No pre-processing is applied and 5-fold cross validation is used, which leaves sample sizes of 22068 in each fold. To evaluate the performance of different parameter values like in the minimal node size, m_{try} or the split rule and to identify

the best performing model out of all combinations, a grid search was performed.

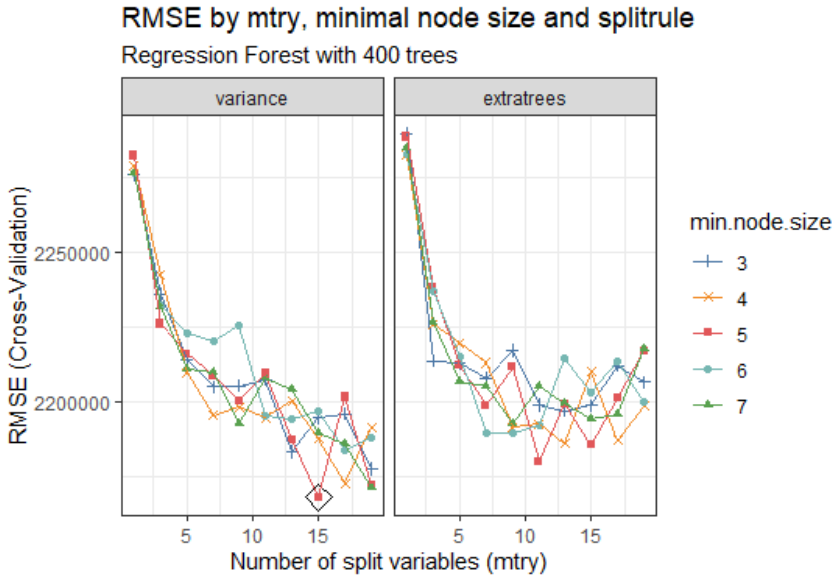

FIGURE 5.3: Performance of all combinations of tuning parameters tested for the simple regression forest

Figure 5.3 is a way of representing the model performance measured based on the RMSE of the different models evaluated in the grid search. As the graphic shows, the variance and extratrees were tested as split rules, minimal node sizes between 3 and 7 and m_{try} values spaced out between 1 and 20 were evaluated. The best combination of hyperparameters according to the error is circled.

Results of the best model

The characteristics of the model which was selected as the best performing model are presented in table 5.1. The Pearson correlation between the between real and predicted error values is 0.217.

m_{try}	*split rule*	*node size*	*RMSE*	R^2	*MAE*
15	variance	5	2167934	0.18365950	138126.8

TABLE 5.1: Best performing model for cross-sectional information

The performance measures are important tools to evaluate which of several comparable models (with the same number of variables, the same underlying observation, etc.) is the best model. However, they don't provide any information which would let us determine if the selected model is a good model for our purposes. Hence, we need to proceed to perform further analysis. Important insights can be gained by comparing the predicted error in turnover with the real value of this error.

To this end, we first compare the distribution of the predicted errors in turnover with the real errors in turnover. This can be done for example with a box plot, or by directly overlaying the density plots of both distributions, like in figure 5.4. Both distribution don't seem to be terribly different, yet, we can see that the predicted values are more widespread and have a less asymmetric distribution.

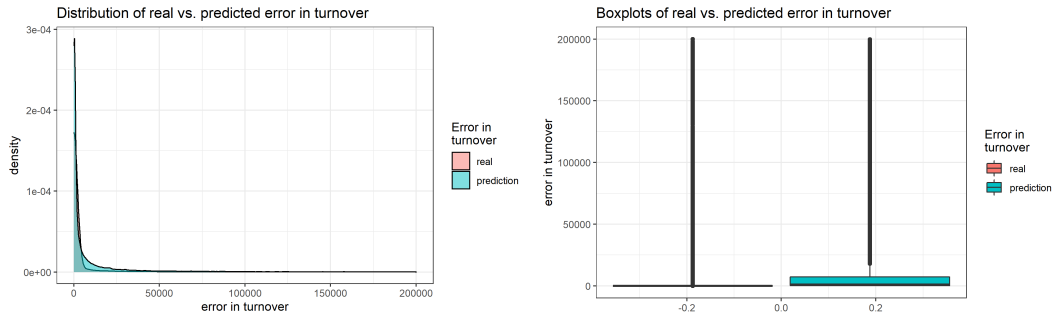

FIGURE 5.4: Comparison between real and predicted error distributions

Another option to evaluate the model is to compare the values in a scatter plot. The scatter plots in figure 5.5 show the different parts of the same plot: The first one shows all the values, while the second one only shows values < 100,000 to illustrate the relation between both variables despite the asymmetrical distribution. The scatterplot does not show a clear correlation between the two variables, in fact the points look rather randomly distributed. It is also noticeable that for many units whose error is actually zero, an error greater than zero is predicted.

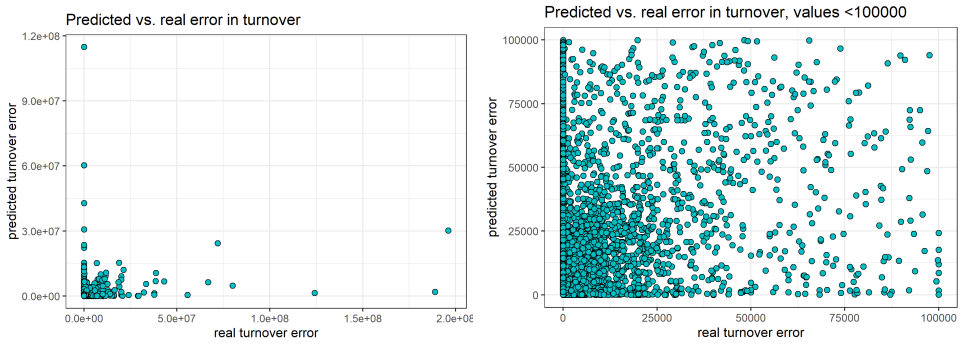

FIGURE 5.5: Relation between real and predicted errors in turnover

This result is in no way satisfactory. As far as the variables used in the model are concerned, figure 5.6 represents the variable importance and shows the raw turnover value as the most important split variable in the forest. The next most important variables, such as the number of employees or the CNAE classification, follow with much lower index values.

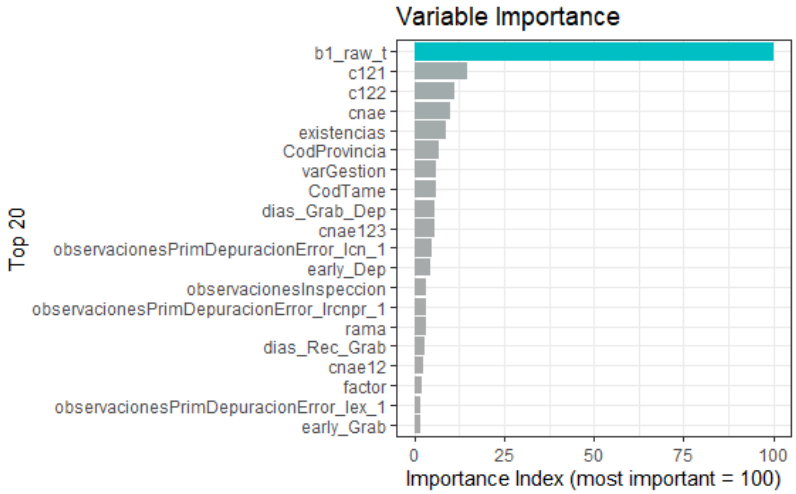

FIGURE 5.6: Importance indices of the most important variables used in the model

However, it must be mentioned in this context that the importance values are not very reliable if the whole model does not work well.

5.1.2 Random Forest based on Cross-Sectional and Longitudinal Information

The performance of the previous results was not as good as it ought to be, which is why in the next step, longitudinal information will be introduced to the model in form of two variables. Using information from the previous month $t-1$, the relative variation of the turnover value $\frac{|Y_t^{raw} - Y_{t-1}^{ed}|}{Y_{t-1}^{ed}}$ will be included in the model, as well as the previous months edited ("true") turnover value itself.

A reasonable number of trees to be build (in this case 300) was obtained in the same way as in the previous section by analysing the error curves in relation to the number of trees. Concerning the number of split variables, it now seems that lower errors are expected for high values of m_{try}[2].

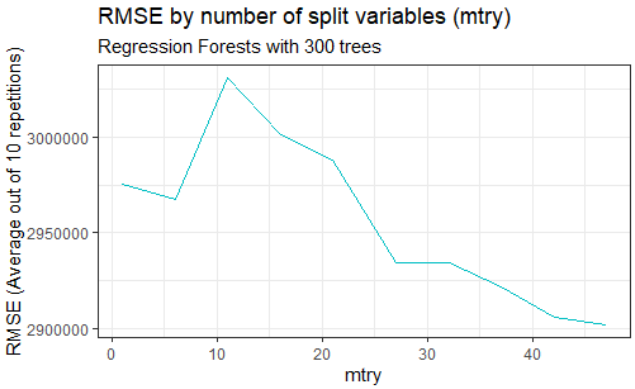

FIGURE 5.7: RMSE by number of split variables (m_{try})

[2]The related graphics can be found in Appendix A

Due to observations that had to be excluded because of missing values, the sample size has decreased a little compared to the models describes in the previous section, the sample size is now n = 24782. In return, the model does now include 47 predictors. Using 5-fold cross-validation, the following combinations of the parameters were created and compared in the resulting models. Like the figure shows, models with variance as split rules performed generally better than those with split rule extratrees. As already expected, the models reach better performance with higher m_{try} values.

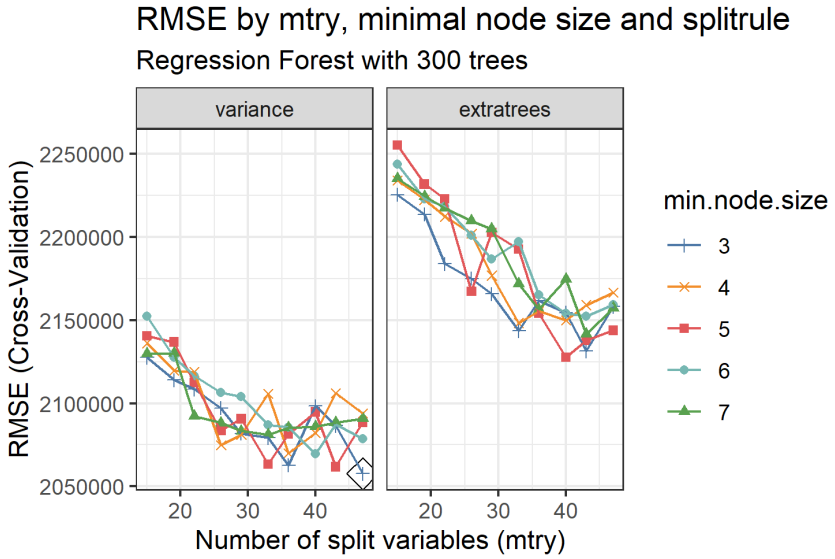

FIGURE 5.8: Performance of all combinations of tuning parameters tested for the simple regression forest with longitudinal variables

Results of the best model

The selected best fit actually considers all 47 variables for each split, which is equivalent to *bagging* and problematic in terms of the correlation between the trees in the forest. Besides, the selected model has a minimal node size of 3, meaning that the model is more complex than a default one, which would be have a minimal node size 5.

m_{try}	*split rule*	*node size*	*RMSE*	R^2	*MAE*
47	variance	3	2057632	0.2691465	95530.74

TABLE 5.2: Best performing model for cross-sectional and longitudinal information

The Pearson correlation between the between real and predicted error values is 0.582, so considerably higher than in our last model with only cross-sectional information. Analysing the density distributions in figure 5.9 both distributions look similar, however the boxplot reveals that again the distribution of predictions is less skewed than the distribution of the real error values.

Looking at the scatterplot in figure 5.10 with plots the datapoints of both the predicted and real error values, a very conspicuous pattern can be observed. It gives the impression that a third quantity interacts with the relationship and splits it in two parts. In the following section we will try to find ways to better represent the characteristics of the underlying data in the forests.

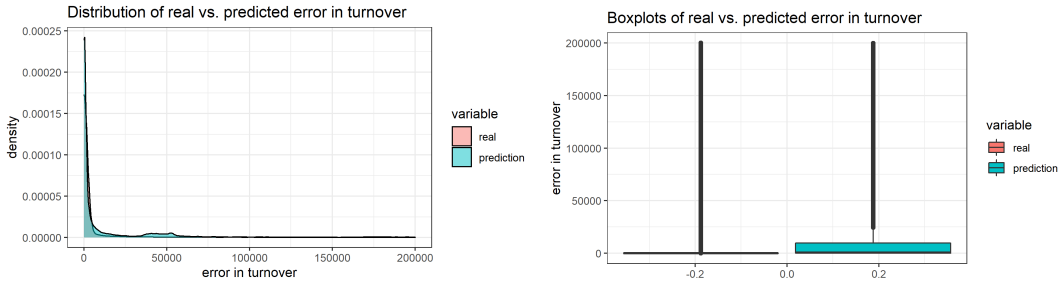

FIGURE 5.9: Comparison between real and predicted error distributions

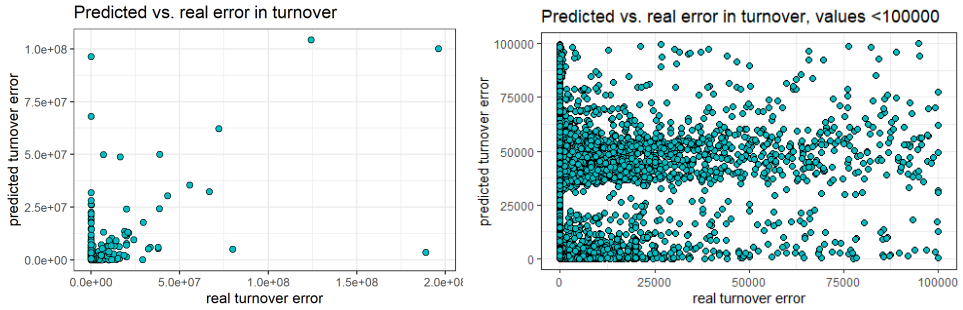

FIGURE 5.10: Relation between real and predicted errors in turnover

Regarding the importance of the variables, the raw turnover value has been displaced from its position as the most important variable by the edited turnover value of the previous month. It is now only in second place but still considerably important. Instead of one variable, as in the previous model, three variables are now of increased importance, since the difference between the turnover value and that of the previous month is also important in the third place. We therefore conclude that the inclusion of the two new variables with longitudinal components was a useful extension of the model.

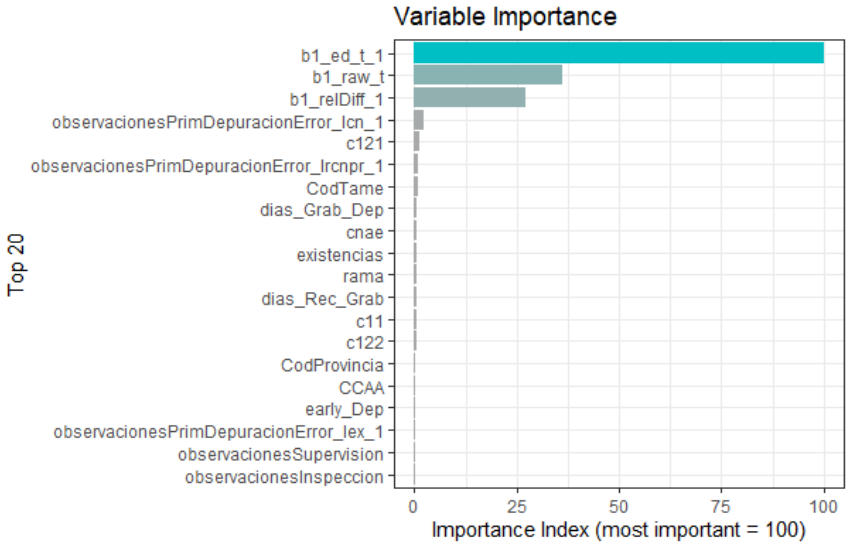

FIGURE 5.11: Importance indices of the most important variables used in the model

5.2 Two-step random forest approach for semi-continuous data

Although random forest is an extremely flexible method with regard to the characteristics of the data and doesn't require strict assumptions about its distribution, the previous approach has not produced the desired results. It seems like there are two main issues that prevent the models from adjusting to data successfully:

1. Observations with missing values in turnover
2. Semi-continuous target variable

The target variable *error in turnover value* has missing values due to missing values in the *raw turnover* variable. Like explained before, the value of the error variable was therefore set to the edited value of *turnover*, which is equivalent to set missing variables in *turnover* to 0. However, the distribution of the *error* for these observations is quite different from the distribution of errors in the rest of the data, which can be seen clearly, in both the density plot and the boxplot in figure 5.12. The need to build a separate forest for these observations was then derived from the findings.

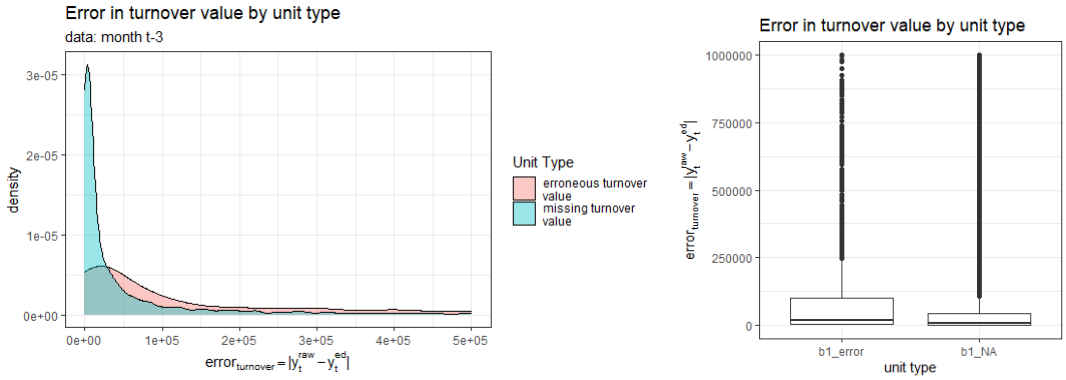

FIGURE 5.12: Distribution of error in turnover by unit types in the training data

But even if we analyse observations with missing values in *turnover* separately, another problem remains: The target variable is not just highly skewed, but can in fact be described as a semi-continuous variable. The vast majority of the observations (in the observed months around 85% - 95%) already contain a correct turnover value in the raw data and therefore have an error equal to 0. Hence, the error variable can be disaggregated into a binary part and a positive-valued continuous part:

$$error_{b1} \in 0 \cup \{LO, .., UP\},$$

where the finite lower bound $LO = 0.01$ and the infinite upper bound $UP = \infty$. To address these issues caused by these characteristics of our data, the alternative two-step approach, that treats the data in three separate parts, is applied. The different forests are built based on the different types of units in the data, see figure 5.13 for a schematic representation.

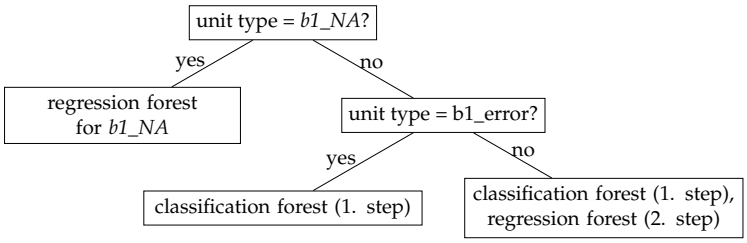

FIGURE 5.13: Decision forests built depending on the unit type in the two-step approach. Own figure.

Splitting up the data into three types of units according to the value in the turnover variable gives rise to a new problem: The data basis is significantly reduced, especially for erroneous observations which represented about 6% in the original training data, and the units with missing values in *turnover*, which represented about 9% of the data. The data base was therefore amplified by data from additional months and the available unit by type are represented in figure 5.14. Like this, $n_{noNA} = 144652$ for non-missing observations, from which $n_{b1error} = 3619$ and $n_{NA} = 9906$ observation with missing value in *turnover* were available to train the three forests.

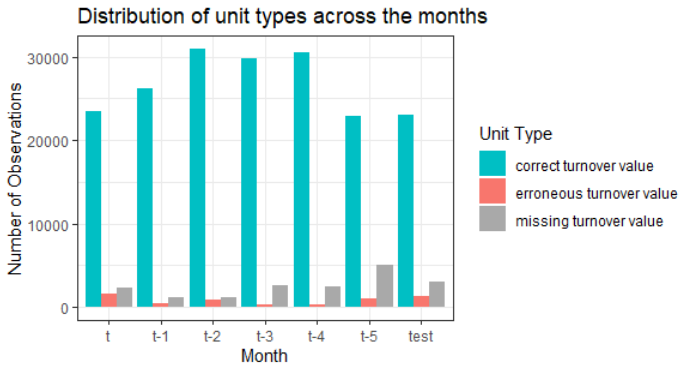

FIGURE 5.14: Distribution of unit types in the training data

5.2.1 Step 1: Classification forest for subset without missing value in turnover

Like mentioned previously the first step of the two step approach aims to classify the observations into those with an error in the turnover error variable (*yes*) and those without (*no*), leaving those observations with a missing value in the turnover variable out of the analysis.

Fitting a simple classification forest to this data with $n = 144652$ using 5-fold cross-validation yields promising results with an overall accuracy of 0.946. However, if we take a closer look at the classification tree that has been built, a fundamental problem becomes apparent. In this context we analyse the table of incorrectly and correctly classified observations, the so-called confusion matrix.

		predicted	
		error	*correct*
true	*error*	92	1269
	correct	47	22955

TABLE 5.3: Confusion matrix of the forest applied to the imbalanced data set

It seems like the classifier worked well for observations with a correct turnover value, as only 47 of 23002 correct observations have been wrongly classified as erroneous observations, which corresponds to a specificity of almost 1. But as we can see, only 92 of the 1361 erroneous observations where detected as such. The true-positive rate, also called sensitivity, is therefore only 0.07, which would be a highly displeasing result because it means that the model doesn't work well to detect erroneous observations as such and thus completely misses its purpose.

The problem is related to the fact that the two classes of our binary classification problem have a highly unequal number of observations, like figure 5.15 shows. A well-known problem in dealing with these imbalanced data is that machine learning algorithms tend to be biased towards the majority class and ignore the minority one. This problem also occurs in random forests, because the model performance is optimized based on overall accuracy. If the observations with an error represent only about 5% of the data. That means that an OOB error as low as 5% is still unacceptable, because an overall accuracy of 95% means that the model is only as good as one that would always predict the majority class.

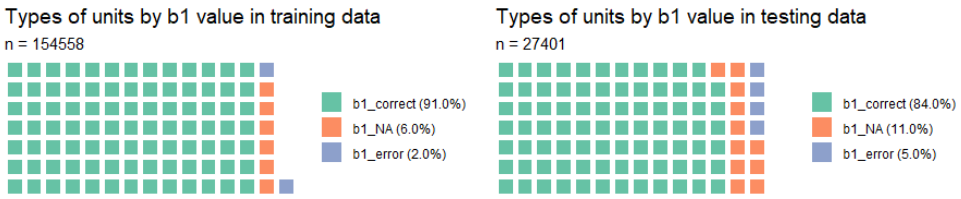

FIGURE 5.15: Distribution of unit types in the training and test data

Methods to handle imbalanced data

To address the problem of imbalanced data, measures can be taken either at the data level or at the algorithm level. The former are called sampling methods and include a number of techniques to balance the imbalanced data (Sonak and Patankar, 2015). One of these sampling methods is *under-sampling* or *downsampling*. In downsampling, a fraction $r \in (0, 1]$ of the majority class (often times, and also in our case, the negative class) is selected, to either diminutive the imbalance or fully level it by selecting a subsample size equal to the size of the minority class. As the training set will be much smaller than the original imbalanced set, this technique is also convenient regarding computational resources. On the downside, it implies a loss of information, which is why the data set should be sufficiently large to consider downsampling.

Oversampling or *Upsampling* allows to balance the classes without having to dismiss any of the available information by replicating units of the minority class type. While random oversampling randomly samples (with replacement) units of the minority class to duplicate them, informative oversampling uses the k nearest neighbours of all the minority samples to create synthetic units. Methods like *SMOTE* (Synthetic Minority Over-sampling Technique) can also combine the two previous techniques: New minority samples are created, while the majority class is sub-sampled to balance the data without loosing to much information but reducing the risk of overfitting due to a highly synthetic data set (Sonak and Patankar, 2015). Using the explained sampling methods, the following data sets were created:

Model	$N_{b1correct}$	$N_{b1error}$
Original data set	141033	3537
Downsampling	3537	3537
SMOTE 1:	14148	10611
SMOTE 2:	7074	7074
SMOTE 3:	10611	10611

TABLE 5.4: Balanced data sets created based on the original data set by applying several sampling methods

The three variants of the SMOTE differ in the proportion of synthetic observations created and in the proportion of the subsampled majority class. As shown in table 5.4, SMOTE 2, for example, is a fully balanced sample, in which the number of minority observations was doubled by creating synthetic units and then the same number of observations was drawn from the majority class.

For each of the created data basis, a modelling process using grid searches was carried out in order to find the optimal tuning parameters for these models. For each kind, the best resulting fit was selected to continue working with it. In the following, model performances of these models based on the different techniques are compared. The model based on the original data set, which was already mentioned before, serves as a reference. A model with upsampling was tested on a subset of the data, but its computation was too time intensive and the results were not good enough to pursue this possibility further. In these models, the ROC (receiver operating characteristic) metric was used instead of accuracy to select the optimal models. A ROC graphic represents the relative trade-offs between benefits (true positives) and costs (false positives) (Fawcett, 2006) and therefore takes the balance of these two measures into account better than a simple accuracy value.

	Original data set	Down-sampling	SMOTE 1	SMOTE 2	SMOTE 3
m_{try}	7	12	19	22	17
node size	10	7	5	5	1
Sensitivity	0.078	**0.824**	0.637	0.767	0.689
Specificity	0.999	0.818	0.933	0.878	0.913
AUC	0.899	0.889	0.896	0.893	0.896
Accuracy	0.946	0.744	0.873	0.807	0.855
Bal. Accuracy	0.533	**0.771**	0.745	0.760	0.758

TABLE 5.5: Original and balanced data sets used to build the classification forest

Table 5.5 shows the selected hyperparameter values and quality measures related to the models. Although the models differ in terms of sample size and model parameters etc., they are comparable in that the best possible variant of each model type is compared here, which could be achieved by optimising the hyperparameters. First of all, it is noticeable that all models achieve quite similar values in the ROC metric. This is also apparent looking at the ROC curves in figures 5.16 and 5.17 shown below. There the curves between the different models are compared and the area under the curve is indicated with "AUC". In a perfect curve, where maximum sensitivity and maximum specificity are possible at the same time, this number would be 1.

We can also observe that the model based on the downsampled data achieves the best values in terms of sensitivity and balanced accuracy. The balanced accuracy

is simply an average of the sensitivity and specificity values. Even if the downsampling model scores the worst in overall accuracy, for our purposes we need to prioritise a model that is good at identifying erroneous values as such. For this reason, and not least because of the computational advantages, the model is selected for further calculations.

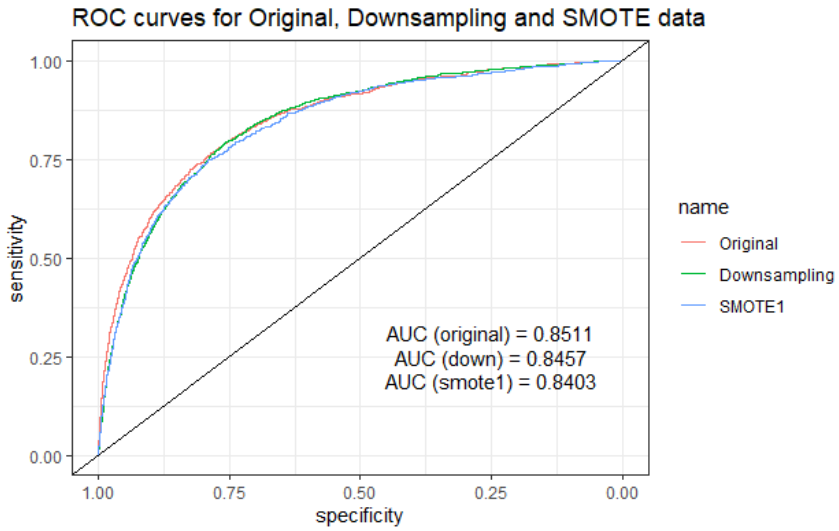

FIGURE 5.16: ROC curves comparing model performance for different sampling methods applied to the original imbalanced data

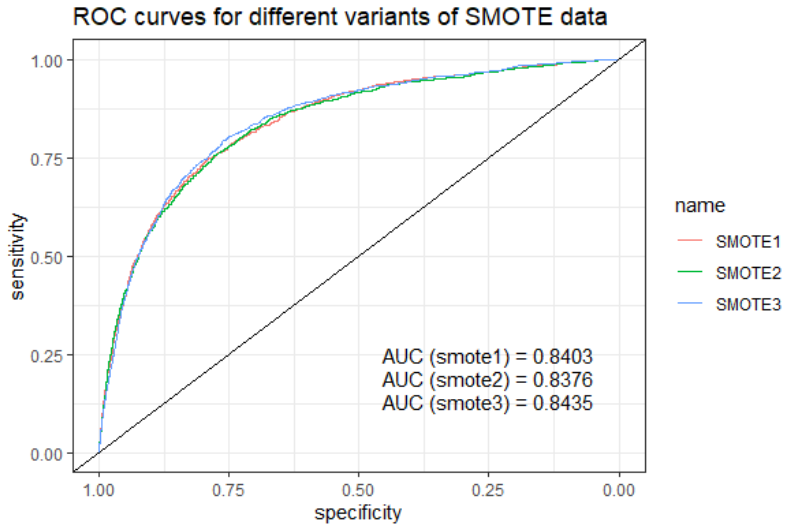

FIGURE 5.17: ROC curves comparing model performance for different SMOTE data

The variable importance values of the selected classification model are distributed as shown in figure 5.18. An interesting observation is that, unlike the regression models in the previous chapter, which were based on all data, it is not a few variables that are decisive, but that the burden of decision making is less unevenly distributed across several variables. In the model, the most important variable to determine whether an observation has an erroneous turnover value, is the relative change in the turnover value compared to the previous month. The second and third variables that appear to be important are whether the unit was revised in the first or second

half of the month and the number of days between the recording of the data and the data editing. Data that took a long time to be processed may be suspected to have a problem and this might be an indication of an incorrect value. A comment from the data collection supervisor also seems to be an important sign. Only after these variables the raw turnover value, the adjusted value of the previous month, the sampling weight and the economic classification (*rama* and *cnae*) appear in the list.

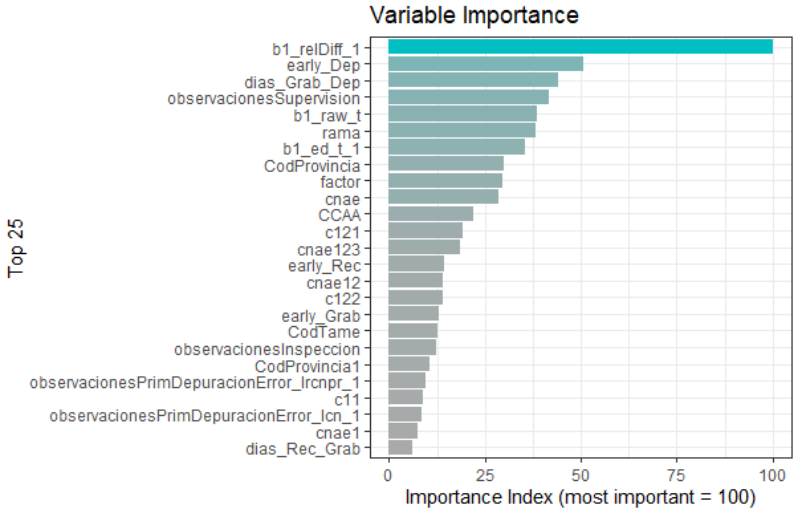

FIGURE 5.18: Importance indices of the most important variables used in the classification model

5.2.2 Step 2: Regression forest for subset without missing value in turnover

After building a classification model as a tool to decide whether an observation has en erroneous value or not, a second modelling step is required to estimate the magnitude of the error for the observations that are found to be erroneous. As splitting up the data resulted in very small training data sets for this second step forest, because there is only a small amount of erroneous values around 200-1500 observation depending on the month, the increased data basis combining data from various month was crucial for this step. An analysis building forest based on the data on singular months, not only showed that this would be an insufficient data basis, but also revealed that different variables seem more or less relevant depending on the month, so the enlarged data base for the training phase could also protect us against overfitting.

The construction of the trees was performed as before in other sections using a grid search, and in this case with bootstrapping with 10 repetitions as validation method. The initial idea was to base the regression forest only on erroneous values as training data, as the quantity to be estimated was the error. For the prediction, however, the forest was applied to the totality of the non-missing observations, as error predictions are needed for the calculation of the score function for all observations, even if the prediction of the error should be close to zero for those observations which were classified as correct by classification forest. In this context, it has been shown that the forest has problems with this and predicts quite high values for many of the zero error observations, as the graphic in figure 5.19 below clearly shows.

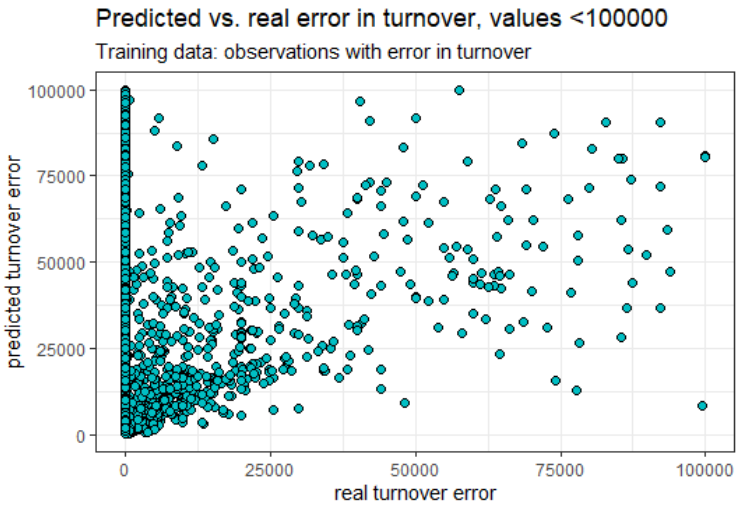

FIGURE 5.19: Scatterplot of real vs predicted error in turnover

This problem may be due to the fact that the regression forest is expected to make predictions on a type of observation that it does not know from the training data. In addition these observations make up the vast majority of the test data (remember figure 5.15). To counteract this problem, a subsample of the non-erroneous observations was added to the database. For this modality, three different compositions were explored, with sample proportions of 50/50, i.e. $s_{b1correct} = n_{b1error}$, $s_{b1correct} = \frac{n_{b1error}}{2}$ and $s_{b1correct} = 2 * n_{b1error}$, of which the first version turned out to be the one that gave the best results.

Although the problem could not be solved completely, the performance of the regression model could still be improved. Figure 5.20 shows the comparison between the density distributions of the real and predicted error values for both the original and selected model (with 50/50 incorrect and correct values). Even if the two density distributions are not congruent in the selected model, they are much more similar than in the previous one. The selected model was built with extratrees, node size of 3 and considering 45 of the 47 predictors at each split. The squared correlation between the real error values of those predicted by the model R^2 is 0.57.

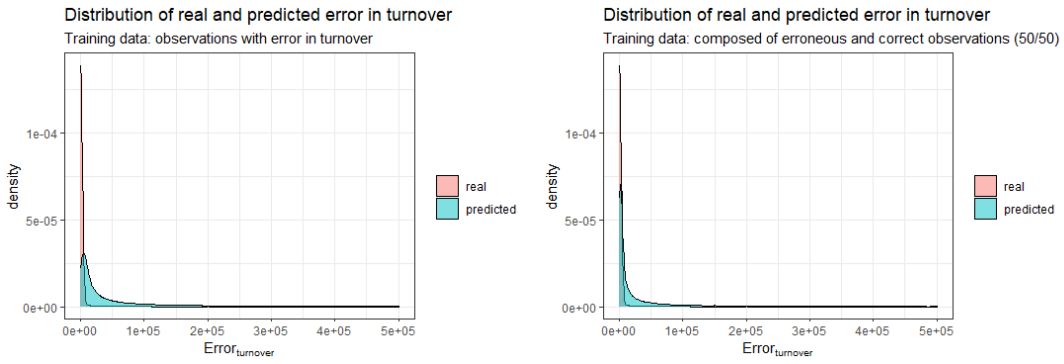

FIGURE 5.20: Distribution of real vs predicted error in turnover

To make the evaluation of such asymmetric data easier, we came up with the solution of ranking the real error values as well as the predictions and displaying the relation between those ranks. The according scatter plot can be found in figure 5.21 and shows a pretty clear correlation between both measures, although there are also

some values further away than desirable from the imaginary line. Overall, it seems that a solution has been developed on which we can build the next steps. Please note that the accumulated observations on the right border with the same rank value result from the fact that the error value is zero for many observations and therefore they are assigned the same rank. In this sense, there are only distinguishable ranks for units with an error > 0.

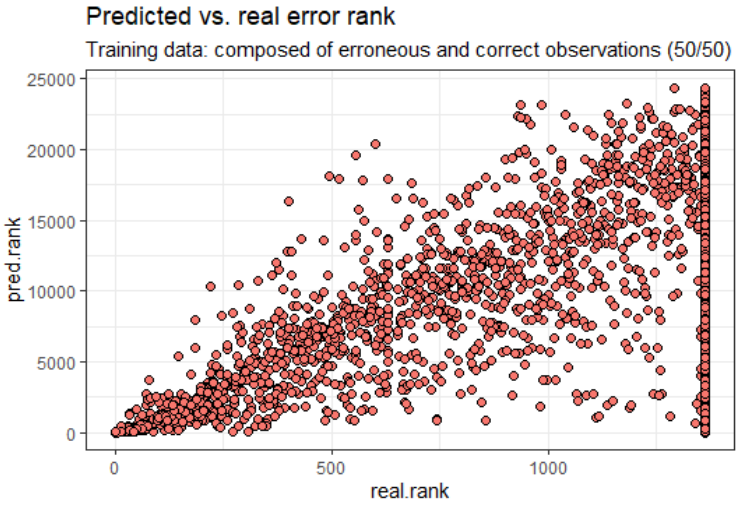

FIGURE 5.21: Scatterplot of real error ranks vs. predicted error ranks for the regression forest

Analysing the importance values of the variables of the model, which are represented on terms of indices in figure 5.22, we conclude that in contrary to the classification tasks, the regression forest is mainly relies on two predominant features, namely the value of the inventory and the raw turnover value. Next important variables would be the edited turnover value from the previous month and relative change from this value to the raw value in the reference month. However, all the following variables have very low importance compared to the first two.

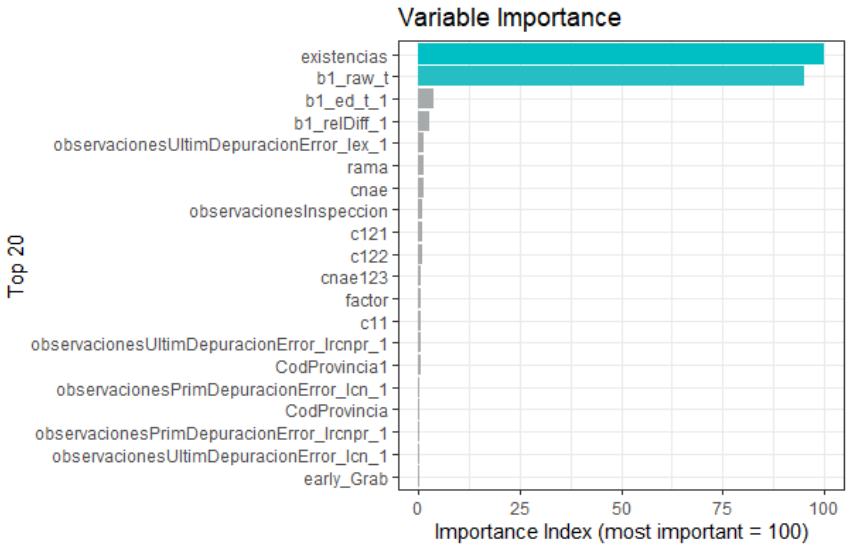

FIGURE 5.22: Importance indices of the most important variables used in the regression model

5.2.3 Regression forest for subset with missing values in the target related variable turnover

In the last step, before we can calculate the local scores for all units, we have to build a regression forest that estimates the error for the observations that have a missing value in the variable turnover. As described above, the target variable in the training data is defined by the value that an expert has finally assigned to each observation.

Based on a sample with size $n_{NA} = 7455$ observations and 45 features we build the forest again using a search grid and in this case 10-fold cross-validation, resulting in the following model as the best fit:

m_{try}	*split rule*	*node size*	*RMSE*	R^2	*MAE*
45	extratrees	4	2154267	0.905	185230

TABLE 5.6: Best performing model for $b1_{NA}$ observations

We can analyse this model based on the measures previously introduced. If we compare the density distribution of the real and predicted error values, it seems that they are quite close to each other, although the predictions, as we have observed before, do not fully reach an asymmetry as extreme as the one of the real values.

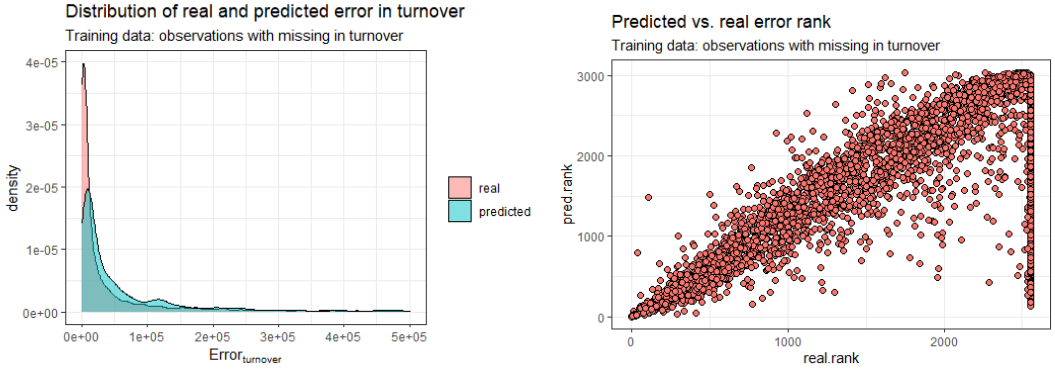

FIGURE 5.23: (a) Density of real vs. predicted errors (b) Scatterplot of real error ranks vs. predicted error ranks

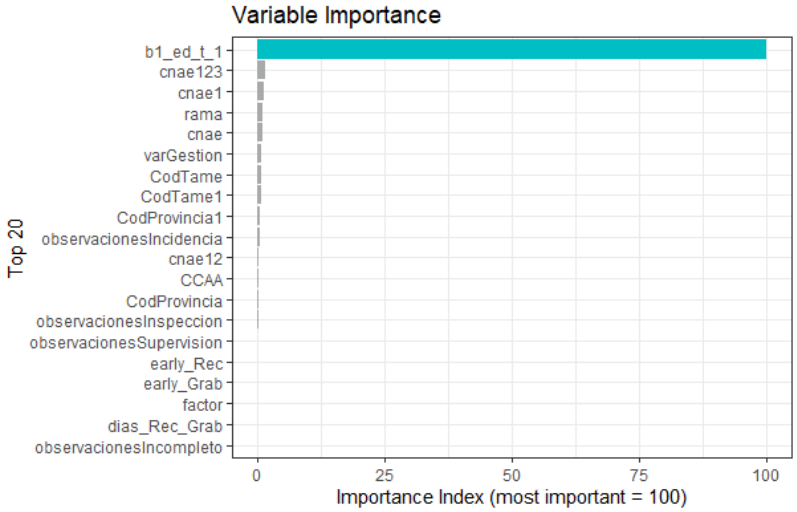

FIGURE 5.24: Importance indices of the most important variables used in the regression model for missing turnover observations

Regarding the variable importances to the variables, the regression forest heavily depends on one predominant variable, the edited turnover value from the previous month of each observation. This is not surprising if one keeps in mind that information about the raw value, which was very important in the previous regression forest, is not available here. The next important variables are related to the economic activity classification variables, but again very far behind the most important variable.

5.3 Evaluation of the results

The first conclusion we can draw, given the large differences in the variables that are relevant in the various models, is that splitting the data into three parts was a sound decision to take. The results of the selected models of this two-step approach can now be combined into unit scores according to the procedure we presented in chapter 2. We obtain score values based on

$$s_k \quad = \quad d_k \cdot \hat{p}_k \cdot \mathbb{E}_{\epsilon_k}, \tag{5.1}$$

with d_k being the sampling weight (*factor*), with $\hat{p}_k$ being the estimated error probability resulting from the selected classification model in the case of observations without missing values and being equal to 1 in the case on observations with missing values, and with $\mathbb{E}_{\epsilon_k}$ being the estimated magnitude of the error resulting from the second step random forest in the case of observations without missing values and from the single regression forest in the case on observations with missing values.

Studying the score values with respect to their quality is a difficult task, since we have no "real" values to check them against. In figure 5.25 below, we are talking about "real" score values, but these simply correspond to $d_k \cdot error$, since there is no equivalent to the error probabilities with which the scores were calculated. If one would rank these scores again, the connection can be shown in the right graphic in figure 5.25.

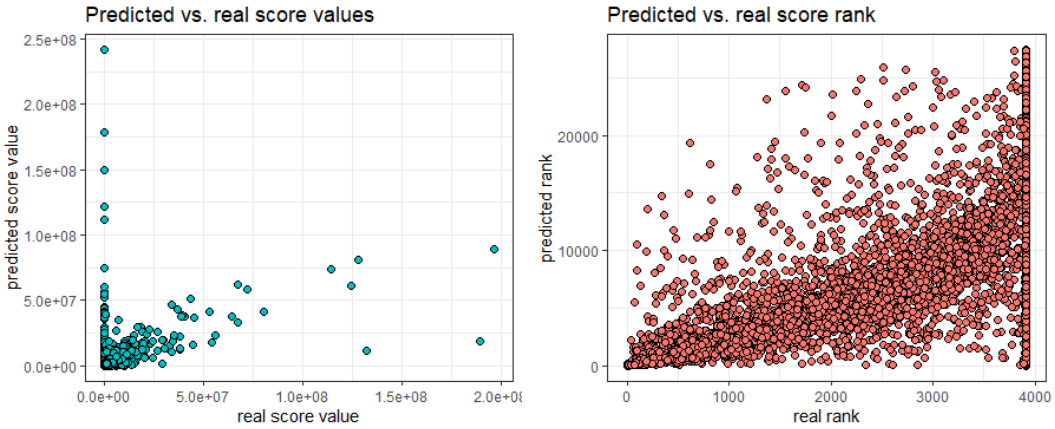

FIGURE 5.25: Scatterplots of real vs predicted score (a) values and (b) ranks in turnover

Selection Efficiency

In order to create a better basis for deciding whether the calculated score values are accurate and serve our purposes, a more sophisticated selection efficiency measure was used, which will be presented in the following. The so-called absolute relative pseudo-bias was also used, for example, in the paper by Arbues et al. (2013) to develop an selection efficiency indicator for score functions. The measure is based on the estimate of the total turnover $\hat{Y}$, which will be used for the disseminated results. $\hat{Y}(n_{ed})$ will denote the estimator that is obtained when n_{ed} items have been selected and edited. The benchmark is $\hat{Y}(n_{ed})$ for $n = N$, which corresponds to the estimator that results after all the data has been edited manually, like it is the current practice in the SSAI. This estimator being denoted $\hat{Y}^0$, the absolute relative pseudo-bias is given by

$$B(\hat{Y}(n_{ed})) = \left| \frac{\hat{Y}(n_{ed}) - \hat{Y}^0}{\hat{Y}^0} \right|.$$

Based on this measure we can now draw the curve of the pseudo-bias evolving according to the number of selected observations. Naturally the pseudo-bias will evolve towards 0, and at the latest with the last observation it will be equal to 0. An averaged random selection of units for editing would correspond to a straight line. An efficient selection of units leads to a monotonously flattening curve, The quicker the curve goes down, the more efficient is our score.

In figure 5.26 the computed curve is presented for our final scores. The graph on the right shows the pseudo-bias based on scores calculated the way it was described here, the graph on the left give a comparison to how the pseudo-bias would evolve if the scores were computed without consideration of the sampling weights, relying only on the estimated error probability and magnitude. As one can see, both curves show decent results, but it is clear that the sampling weights should definitely be included in order to obtain a smooth result and select the units as efficiently as possible. As the theoretical background of selective editing suggests, it is not only the probability of an error that is relevant, but also the impact the error would have on the final estimates, which is inevitably linked to the sampling weights.

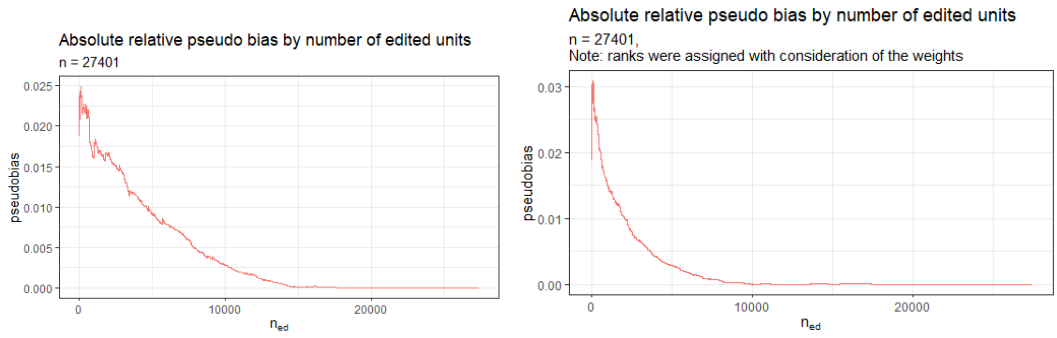

FIGURE 5.26: Absolute relative pseudo-bias by number of edited units, (a) without and (b) with consideration of sampling weights

Figure 5.27 shows that if it wasn't for the observations with missing values in turnover, the selection score would be even more efficient and again shows the importance of the sampling weights for the construction of the score.

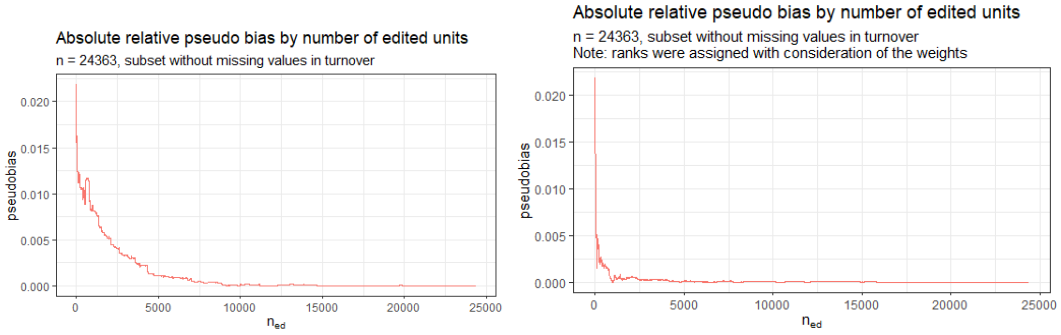

FIGURE 5.27: Absolute relative pseudo-bias by number of edited units for subset without missings

Overall the results are very satisfactory. The fast falling curve shows that the score prioritises the units well. The selection based on the score allows that only the editing of the first 5000 of the approximately 27000 units is necessary to obtain a final estimate that differs only about 0.3% from the value that would be obtained if all turnover values were checked manually.

So far, the efficiency of the score was evaluated for $\hat{Y}$ at the highest aggregation level. In order to deepen the evaluation of the results beyond the global level, the development of the pseudo-bias is also analysed for subsets, namely the autonomous communities and the economic activity classification, which are relevant for the calculation of the final indices. For this purpose, the pseudo-bias for subgroups was analysed in two different ways. The first method would be to re-rank the observations within a subgroup based on the global score ranks which previously assigned to the them. Thus, by re-ordered within the autonomous communities, and respectively within the activity classification groups, a ranking of ranks is created. The second way would be to maintain the global ranks but just draw the graph considering only observations of each subgroup.

The re-ranking by subgroup method allows for an analysis of how the pseudo-bias would fall in this autonomous community or branch if all the editing resources were concentrated to this autonomous community or branch. On the contrary, subgroup specific curves based on the global rank show the behaviour of the pseudo-bias in each autonomous communities or branch if the resources are distributed equally, only in order of global ranking. In some groups this produces pronounced staircase functions, as the subgroup-related pseudo-bias remain the same until the editing resources are directed back to this subgroup.

Looking at the pattern of the pseudo-biases shown in details in the figures in appendix A, it can generally be said that the prioritization of the units can also be successfully applied at these lower aggregation levels, in different autonomous communities and the economic branches.

Chapter 6

Conclusions and Future Work

The aim of this thesis was to explore the applicability of random forests for the selective editing of official data, more precisely numerical variables. In this context, the specific conditions under which National Statistical Offices carry out their work were discussed and it was found that they are currently faced with the challenge of meeting requirements in different quality dimensions. It was also identified that international co-operation, standardisation and innovative methods can be a mean to reduce the resulting quality dimension conflicts. Therefore, previous applications of machine learning methods in statistical authorities were analysed and random forests were recognised as a versatile technique for pattern recognition. The characteristics and functioning of random forests were described, and two concrete approaches were developed about how they can be applied for selective editing using short-term business statistic SSAI as an example. An algorithm for the calculation of a score function could be provided, which efficiently selects influential observations for further manual editing, according to the estimated probability of an error in the turnover value and the influence of such an error. It was found that the selection approach is also applicable to lower aggregation levels.

If influential units are selected efficiently, in a way that only the relevant ones need to be recontacted, then this is not only a benefit in terms of saving manual resources, but also implies that these resources can be reallocated elsewhere, possibly to improve the quality of statistical products in other areas. Decisions are also made more consistent if there is a uniform basis for them and no bias arises from different data managers. The response burden for companies is also reduced when fewer manual recontacts are required, promoting the compliance with the EU CoP P9 principle of non-excessive burden on respondents.

To identify specific points that could be pursued further, we want to mention some aspects that might increase model performance. The missing data in this work was manually imputed, however, there are many options of imputing the data, based on the proximity measures inherent to the random forests built or other algorithms like knn algorithms just to mention two examples. More longitudinal data could complement the model if it would be combined with time series models that make use of past months information. Concerning the handling of imbalanced data, non-sampling methods like cost-sensitive learning could be tested. Also, the approach has so far been tested only for one item, the *turnover* value. A future work could apply it to other target variables such as the value of the inventory or the number of staff. When different item scores have been developed, a unit score function can be applied to combine these, like explained in chapter 2. From a methodological point of view, semi-continuous variables as target variables in random forests are a field that does not yet seem to be very well studied and might be worth exploring in greater detail.

The long-term goal of a work like this one is of course to make the tested methods applicable to different kinds of similar data and for other statistical institutions, to offer new perspectives and combine them with related techniques. As metadata information about the production process increments, so will the variables eligible for a predictive model. In the presented case, available features about the production process were manually revised to preselect promising variables for the model. Instead of manually reconstructing the underlying data basis every time new information is available, future implementations should revise possibilities of connecting the modelling process to the data infrastructure available to automatically consider and preselect model features out of all the available information.

Our work concentrates on the selection of suspicious units. Some of the existing methods and tools for editing and imputation with a focus on machine learning methods, like CANCEIS from Statistics Canada or HoloClean, which was developed by the Standford University, combine the detection of error in the editing phase directly with the following imputation of erroneous values. Like this, recontacts are completely avoided (Lange, 2020). Based on the findings of our work, we want to think about ways to integrate the approach we have developed into more broadly oriented software solutions for the E&I process. An advantage of the application of random forests in the way it was done here, is that no manually predefined edits are required for the detection of suspicions values or units. The points mentioned in the last two paragraphs call for the development of a flexible tool, which can be standardized in the spirit of the CSPA, in accordance to the EU CoP P1b principle of coordination and cooperation of the NSIs.

As Hedlin (2003) accurately points out, instead of searching ways to best detect errors, in general it would be obviously better to pretend errors from happening. Trying to fix them later has its costs and limitations, and to allocate a huge part of the institutions resources to editing is not a desirable situation. Some errors, e.g. so-called inliers are usually not even detectable with micro editing. This leads us to an example of a graphic from this work that wasn't mentioned before but illustrates the problem of undiscovered errors. Figure 6.1 shows the distribution of the relative change in turnover by unit type from the previous month to the reference month. It stands out that under those observations that where found to have an erroneous turnover value, relative difference values around 1 are very common, which then will be corrected in the editing process. However, in the values that were found to be correct, a small accumulation around 1 is still noticeable. We therefore must ask ourselves, how can we know if those are indeed correct values, or if they might rather be errors that weren't detected in the manual editing process?

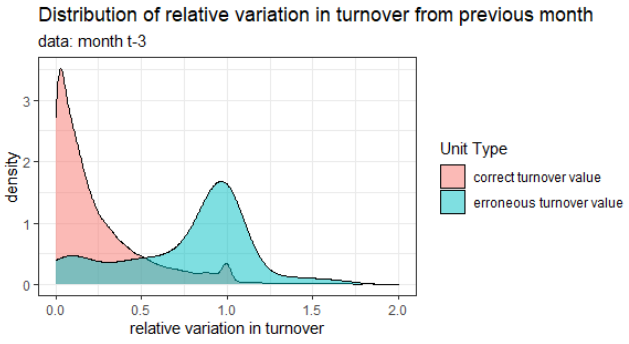

FIGURE 6.1: Density distribution of the relative change in turnover by unit type

Whenever we have described values as "real" values before, we naturally have no chance to know if they actually are. Real values in the context of this work are only the values assigned after manual editing, which are (hopefully) closer to the unknown real value. In an approach like this the predictions of the random forests can only be as good as the manual editing work was. This problem connects to a wider discussion related to the "ground truth", which raises the question of how to ensure a basic minimum of quality in training machine learning methods and how to prevent undiscovered bias from being carried along when manual processes are replaced by algorithms. Another critical point related to the application of machine learning methods could be the communication with the users, if methodological descriptions require an advanced basic knowledge of mathematics or computer science, the way in which results have been obtained may appear like a black box to them. These are issues that need to be addressed in the future, as increasing data volume and sources make the use of innovative methods essential for the development of statistical agencies.

In view of the rapidly changing circumstances of NSIs discussed at the beginning, be it through technical innovation, socio-political circumstances or crises, it should be considered to what extent adaptability and methodological innovation can be structurally anchored in NSIs. The bureaucratic outlines of state institutions must be recognised as a challenge, but at the same time they must be seen as an opportunity, since ultimately bureaucracy, innovation and flexibility are all aimed at the same purpose: providing high-quality information and making it democratically accessible. In order to be able to do justice to this purpose, everything speaks for the idea that quality-based innovation can be driven forward best and fastest in European and interdisciplinary cooperation.

Bibliography

Arbues, Ignacio, Pedro Revilla, and David Salgado (2013). "An optimization approach to selective editing". In: *Journal of Official Statistics* 29.4, pp. 489–510.

Barber, David (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Beck, Martin, Florian Dumpert, and Joerg Feuerhake (2018). "Machine Learning in Official Statistics". In: *arXiv preprint arXiv:1812.10422*.

Biamonte, Jacob et al. (2017). "Quantum machine learning". In: *Nature* 549.7671, pp. 195–202.

Biemer, Paul P. (2010). "Total survey error: Design, implementation, and evaluation". In: *Public Opinion Quarterly* 74.5, pp. 817–848.

Boehmke, Brad and Brandon M. Greenwell (2019). *Hands-On Machine Learning with R*. Available at: `https://bradleyboehmke.github.io/HOML/process.html`, (accessed August 2020). CRC Press.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Coccia, Mario (2009). "Research performance and bureaucracy within public research labs". In: *Scientometrics* 79.1, pp. 93–107.

Crow, Michael M. and Barry L. Bozeman (1989). "Bureaucratization in the laboratory". In: *Research Technology Management* 32.5, p. 30.

Cutler, Adele, David Cutler, and John Stevens (Jan. 2011). "Random Forests". In: vol. 45, pp. 157–176.

De Waal, Ton (Dec. 2013). "Selective Editing: A Quest for Efficiency and Data Quality". In: *Journal of official statistics* 29, pp. 473–488.

De Waal, Ton, Jeroen Pannekoek, and Sander Scholtus (2011). *Handbook of statistical data editing and imputation*. Vol. 563. John Wiley & Sons.

Di Zio, Marco and Ugo Guarnera (2013). "A contamination model for selective editing". In: *Journal of Official Statistics* 29.4, pp. 539–555.

European Statistical System Committee (2019). *Quality Assurance Framework of the European Statistical System (ESS QAF)*. Available at: `https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646`.

European Union (2009). "Regulation (EC) No. 223/2009 of the European Earliament and of the Council on European Statistics". In: *Official Journal of the European Union* 284. amended by Regulation (EU) 2015/759, available at: `https://eur-lex.europa.eu/legal-content/en/TXT/PDF/?uri=CELEX:02009R0223-20150608&from=EN`, p. 1.

Eurostat (2017). "European Statistics Code of Practice". In: *Adopted by the European Statistical System Committee*. available at: `https://ec.europa.eu/eurostat/documents/4031688/8971242/KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7`.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

Granquist, Leopold (1997). "The new view on editing". In: *International Statistical Review* 65.3, pp. 381–387.

Groves, Robert M. and Lars Lyberg (2010). "Total survey error: Past, present, and future". In: *Public opinion quarterly* 74.5, pp. 849–879.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hedlin, Dan (2003). "Score functions to reduce business survey editing at the UK office for national statistics". In: *Journal of Official Statistics* 19.2, pp. 177–200.

— (2008). "Local and global score functions in selective editing". In: *Proceedings of UN/ECE Work Session on Statistical Data Editing* 21-23 April, Vienna. Available at: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2008/04/sde/wp.31.e.pdf.

Ho, Tin Kam (1995). "Random decision forests". In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, pp. 278–282.

James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Julien, Claude (2019). "Progress Report. Background document on the HLG-MOS Machine Learning Project". In: Available at: https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home, (accessed August 2020).

Kim, Seoyong, Wanki Paik, and Cheouljoo Lee (2014). "Does bureaucracy facilitate the effect of information technology (IT)?" In: *International Review of Public Administration* 19.3, pp. 219–237.

Kuhn, Max and Kjell Johnson (2013). *Applied predictive modeling*. Vol. 26. Springer.

Lange, Kerstin (2020). "Automation of E&I processes. Working Paper. Workshop on Statistical Data Editing 2020". In: Available at: https://statswiki.unece.org/download/attachments/282329136/SDE2020_T4_Germany_Lange_Paper.pdf?version=1&modificationDate=1596798047993&api=v2, (accessed August 2020).

LFEP (1989). *Law 12/1989 of 9 May 1989 on the Public Statistical Services*. BOE n. 112, 11 May 1989.

Liaw, Andy and Matthew Wiener (2002). "Classification and Regression by randomForest". In: *R News 2* 3, pp. 18–22.

Ljones, Olav (2011). "Independence and ethical issues for modern use of administrative data in official statistics". In: *Statistical Journal of the IAOS* 27.1, 2, pp. 25–29.

López-Ureña, R. et al. (2014). "Application of the optimization approach to selective editing in the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey". In: *INE Statistics Spain, Working Papers* 4.

Louppe, Gilles (2014). "Understanding random forests". In: *Cornell University Library*.

Luzi, O. et al. (2007). *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys (EDIMBUS), ISTAT, CBS, SFSO, Eurostat*. Available at: https://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf.

MacFeely, Steve (2016). "The continuing evolution of official statistics: Some challenges and opportunities". In: *Journal of Official Statistics* 32.4, pp. 789–810.

Measure, Alexander (2017). "Deep neural networks for worker injury autocoding". In: Available at: https://www.bls.gov/iif/deep-neural-networks.pdf, (accessed August 2020).

Moisen, GG (2008). "Classification and regression trees". In: *In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582-588.*, pp. 582–588.

Molnar, Christoph (2020). *Interpretable Machine Learning*. Lulu.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press.

Olsen, Johan P. (2008). "The ups and downs of bureaucratic organization". In: *Annu. Rev. Polit. Sci.* 11, pp. 13–37.

Pannekoek, Jeroen, Sander Scholtus, and Mark Van der Loo (2013). "Automated and manual data editing: a view on process design and methodology". In: *Journal of Official Statistics* 29.4, pp. 511–537.

Probst, Philipp, Marvin N Wright, and Anne-Laure Boulesteix (2019). "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3, e1301.

Rama, Silvia and David Salgado (2014). "Standardising the editing phase at Statistics Spain: a little step beyond EDIMBUS". In: *INE Statistics Spain, Working Papers* 5.

Revilla, Pedro and Asunción Piñán (2012). "Implementing a Quality Assurance Framework based on the Code of Practice at the National Statistical Institute of Spain". In: *INE Statistics Spain, Working Papers* 4.

Sæbø, Hans Viggo and Anders Holmberg (2019). "Beyond code of practice: New quality challenges in official statistics". In: *Statistical Journal of the IAOS* 35.2, pp. 171–178.

Scholtus, S., R. van de Laar, and L. Willenborg (2014). *The memobust handbook on methodology for modern business statistics (MEMOBUST Handbook)*.

Sonak, Apurva and R.A. Patankar (2015). "A survey on methods to handle imbalance dataset". In: *Int. J. Comput. Sci. Mobile Comput* 4.11, pp. 338–343.

Spain (1978). *Spanish Constitution*. BOE n. 311, 29 December 1978.

Spies, Lydia and Kerstin Lange (2018). "Implementation of artificial intelligence and machine learning methods within the Federal Statistical Office of Germany. Working Paper. Workshop on Statistical Data Editing 2018". In: Available at: `https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Germany_LANGE_Paper.pdf`, (accessed August 2020).

Statistics Spain (2019). *Standardised Methodological Report. Services Sector Activity Indicators (SSAI). Base 2015*. Available at: `https://www.ine.es/dynt3/metadatos/en/RespuestaDatos.html?oe=30183`, (accessed August 2020).

— (2020). *Services Sector Activity Indicators (SSAI). Base 2015*. Available at: `https://www.ine.es/dyngs/INEbase/en/operacion.htm?c=Estadistica_C&cid=1254736176863&menu=ultiDatos&idp=1254735576778`, (accessed August 2020).

Stats NZ (2019). *Data sources, editing, and imputation for the 2018 Census*. Available at: `https://www.stats.govt.nz/assets/Uploads/Methods/Data-sources-editing-and-imputation-in-the-2018-Census/Data-sources-editing-and-imputation-in-the-2018-census.pdf`, (accessed August 2020).

United Nations Economic Commission for Europe (2019a). *Generic Statistical Business Process Model (GSBPM)*. Version 5.1.

— (2019b). *Generic Statistical Data Editing Model (GSDEM)*. Version 2.0.

Vale, Steven (2014). "The Common Statistical Production Architecture: An Important New Tool for Standardisation". In:

Weber, Max (1978). *Economy and society: An outline of interpretive sociology*. Vol. 1. University of California Press.

Wright, Marvin N and Andreas Ziegler (2015). "ranger: A fast implementation of random forests for high dimensional data in C++ and R". In: *arXiv preprint*.

Appendix A

Appendix

A.1 Performance measures by Number of Trees

The following figures show the tree curves that where used to analyse the necessary number of trees for the forests.

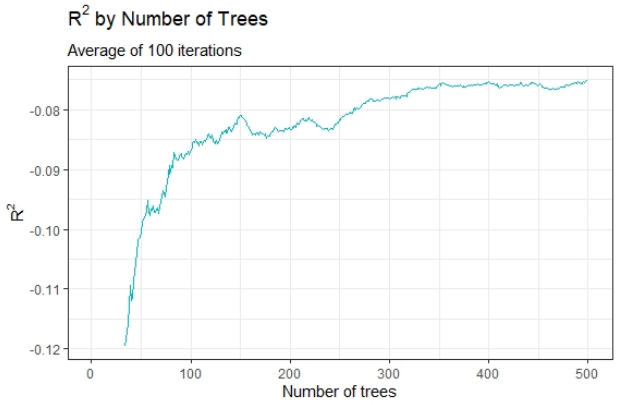

FIGURE A.1: R^2 by number of trees for the simple regression tree

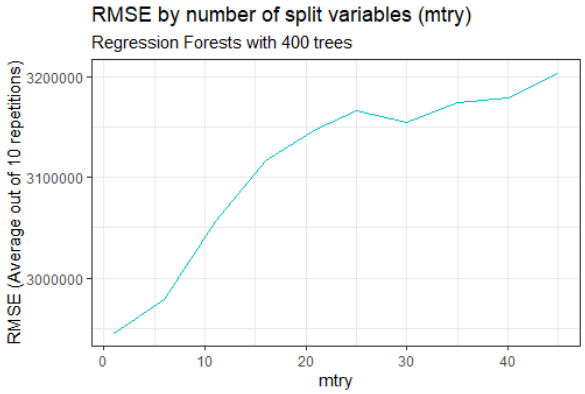

FIGURE A.2: RMSE by number of m_{try} for the simple regression tree

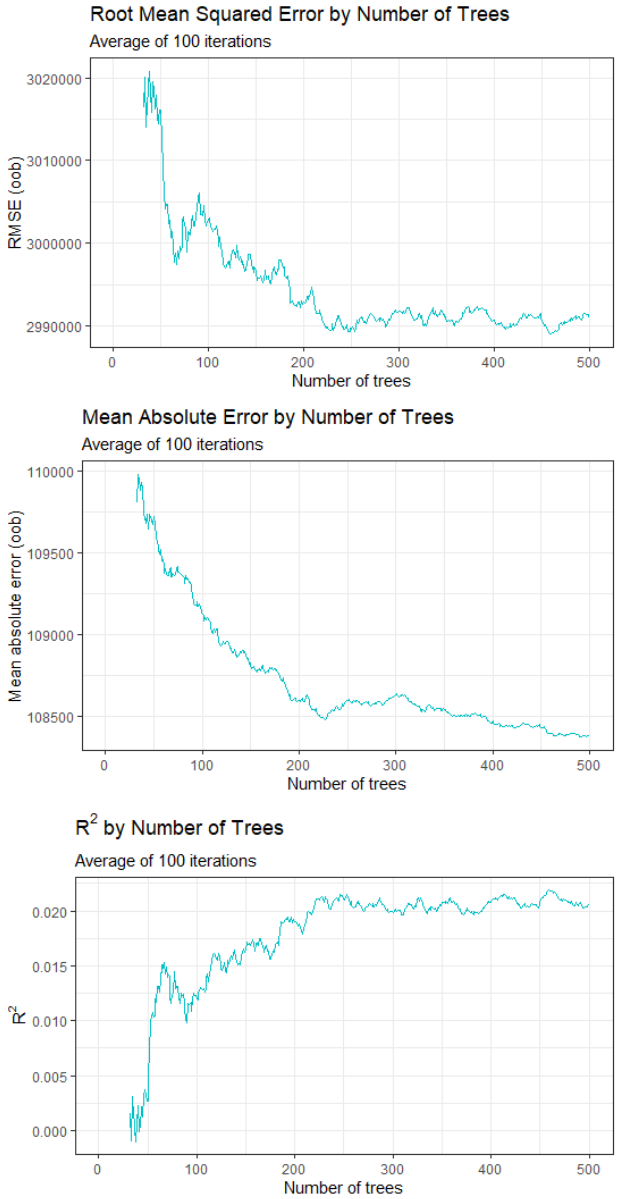

FIGURE A.3: RMSE, MAE and R^2 by number of trees for the simple regression tree with longitudinal variables

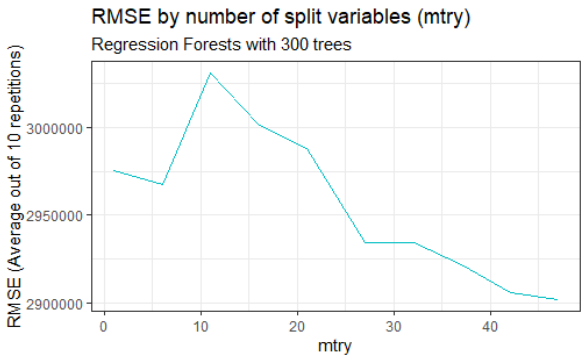

FIGURE A.4: RMSE by number of m_{try} for the simple regression tree with longitudinal variables

A.2 Grid Search Results

To search for the optimal tuning parameter values, grid searches were carried out for all the different types of forests, whose results are represented in the following figures.

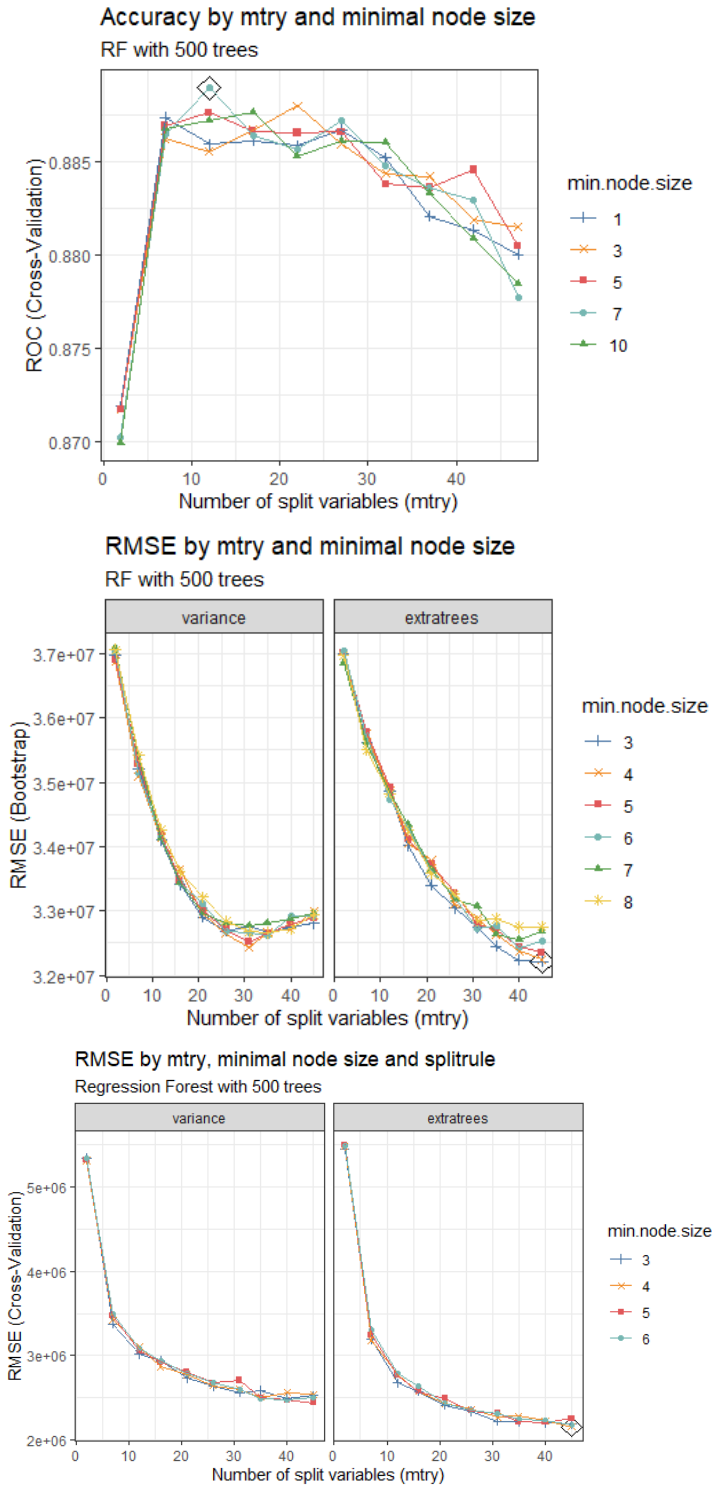

FIGURE A.5: Performance of all combinations for tuning parameters tested for (a) the classification forest, (b) the regression forest and (c) the regression forest for $b1_{NA}$ observations

A.3 ROC Analysis for Classification Forests

The different models that were analysed for each type of decision forest are linked to different results regarding the variable importance, which can be compared in the following figures.

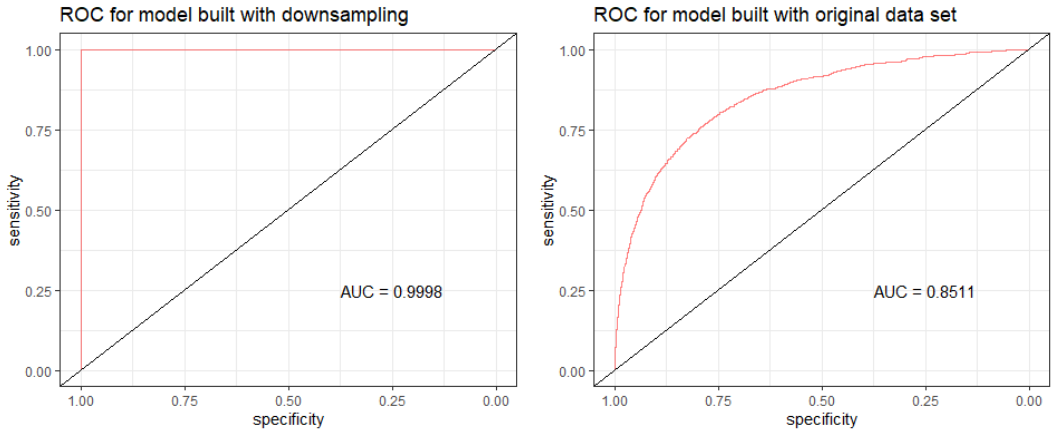

FIGURE A.6: ROCs of model built with original data, (a) train (b) test

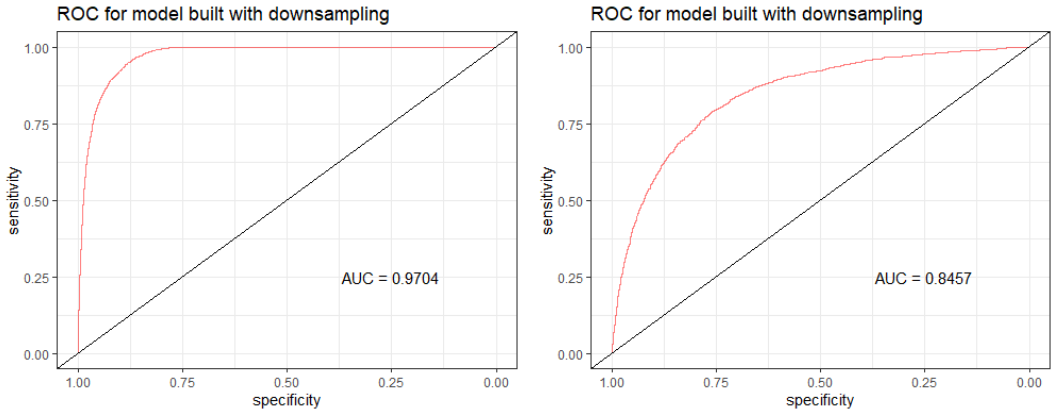

FIGURE A.7: ROCs of model built with downsampled data, (a) train (b) test

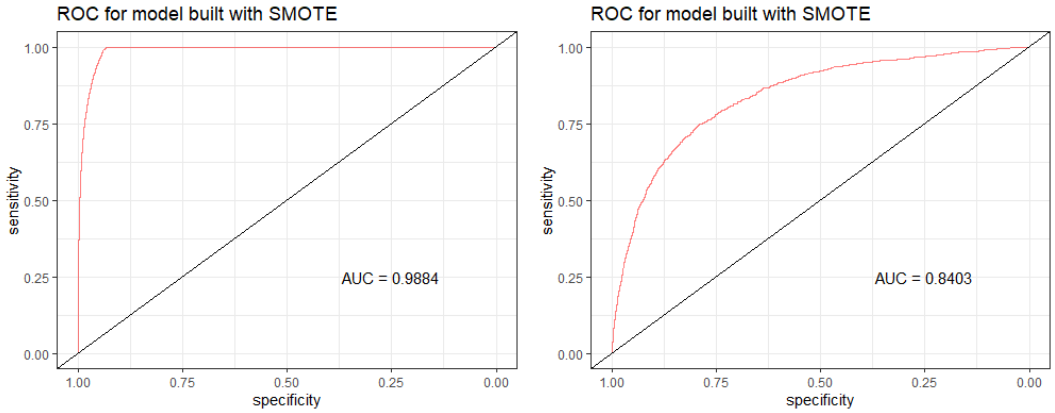

FIGURE A.8: ROCs of model built with SMOTE 1 data, (a) train (b) test

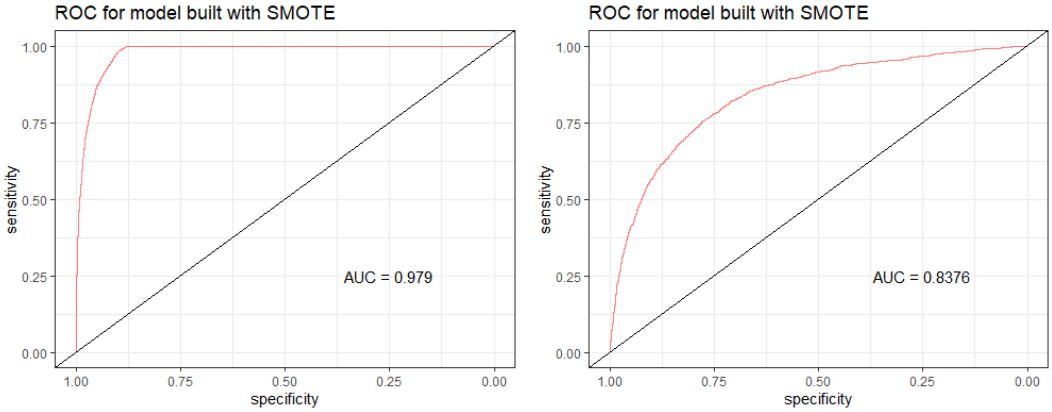

FIGURE A.9: ROCs of model built with SMOTE 2 data, (a) train (b) test

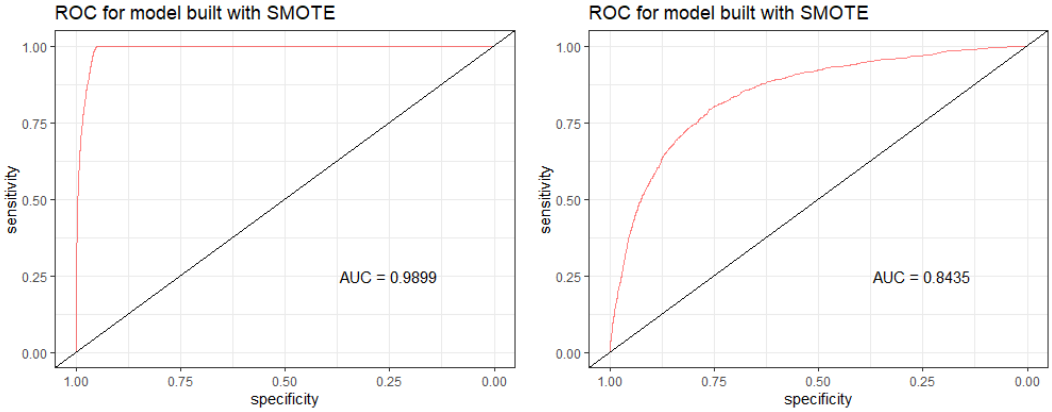

FIGURE A.10: ROCs of model built with SMOTE 3 data, (a) train (b) test

A.4 Variable Importance

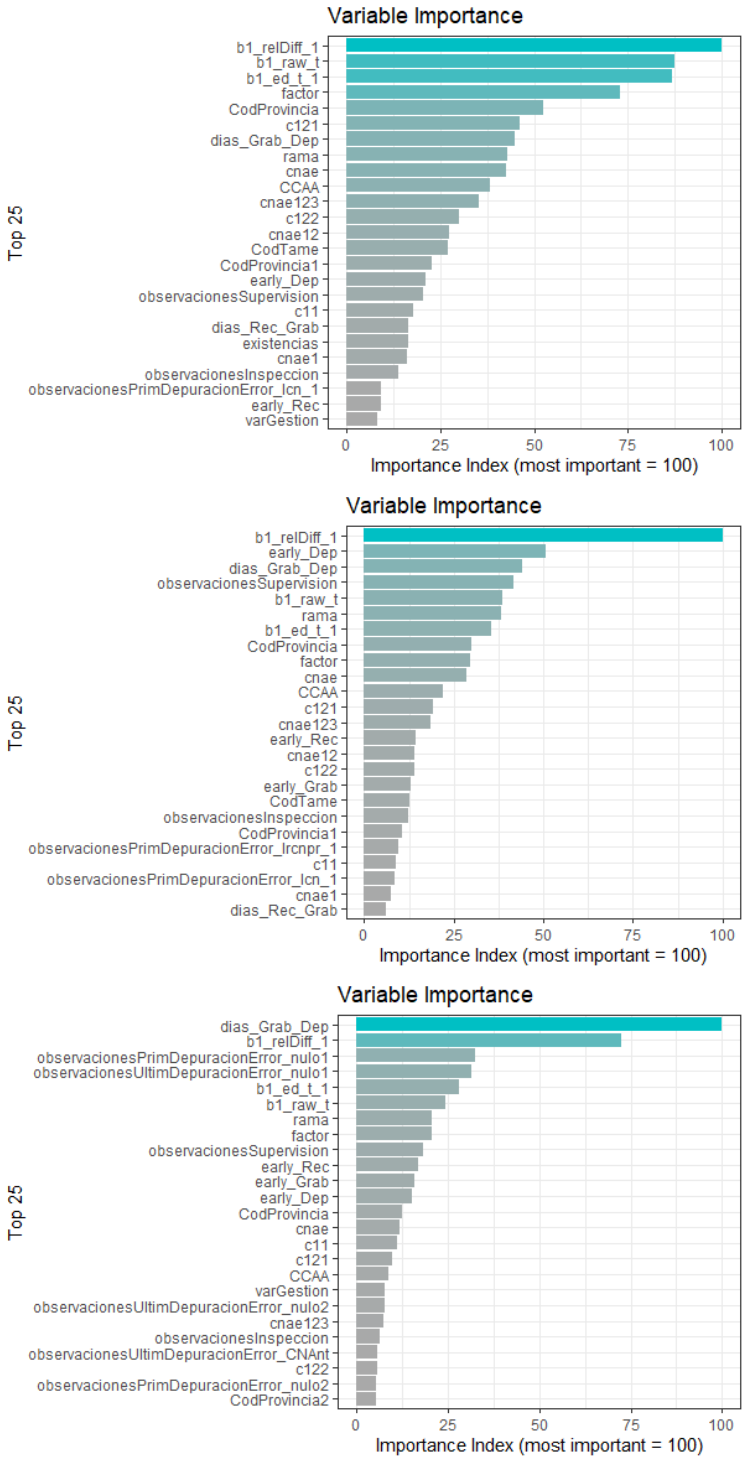

FIGURE A.11: Importance indices of the most important variables used in different analysed classification models (a) original data, (b) downsampled data, (c) SMOTE 1 data)

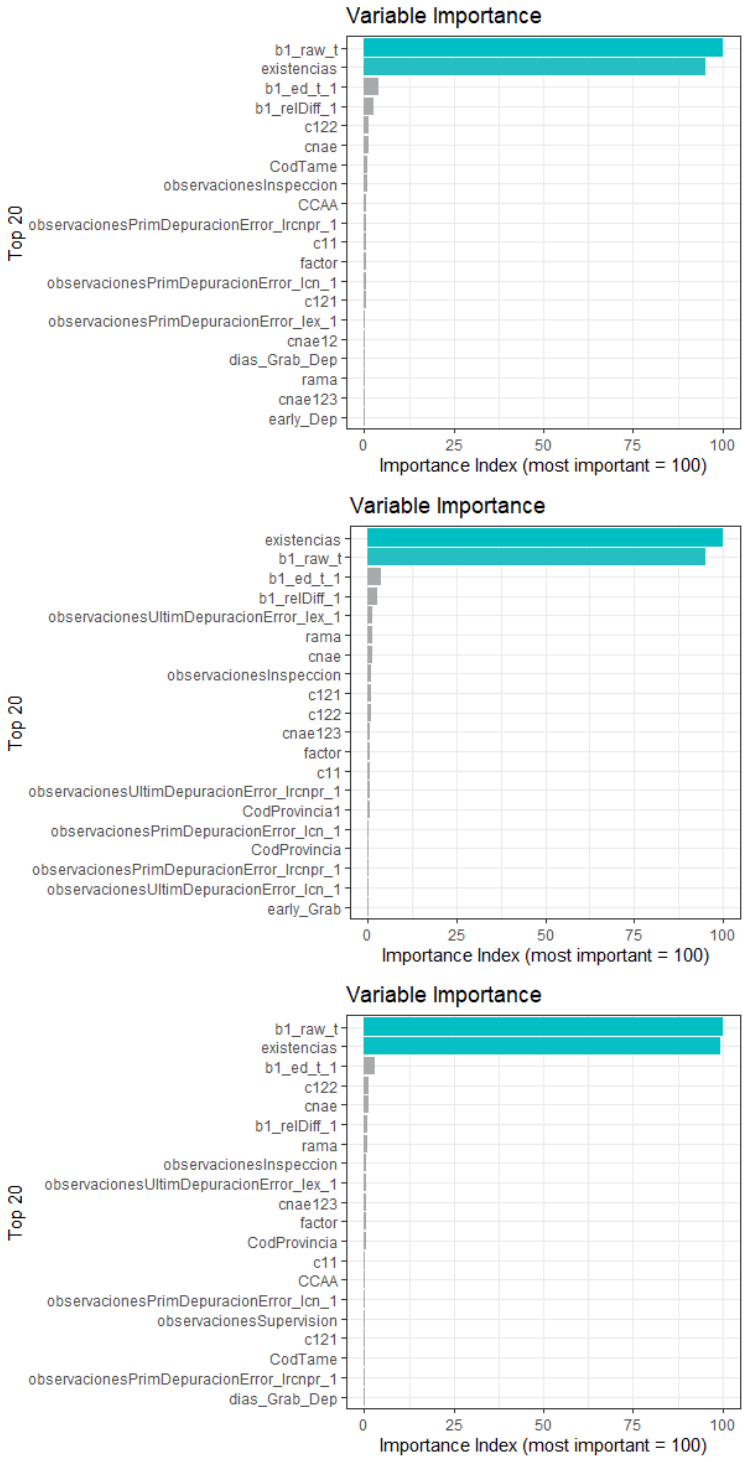

FIGURE A.12: Importance indices of the most important variables used in different analysed regression models (a) without non-erroneous values, (b) with 1/1 erroneous and non-erroneous values, (c) with 2/1 erroneous and non-erroneous values

A.5 Analysis of Selection Efficiency

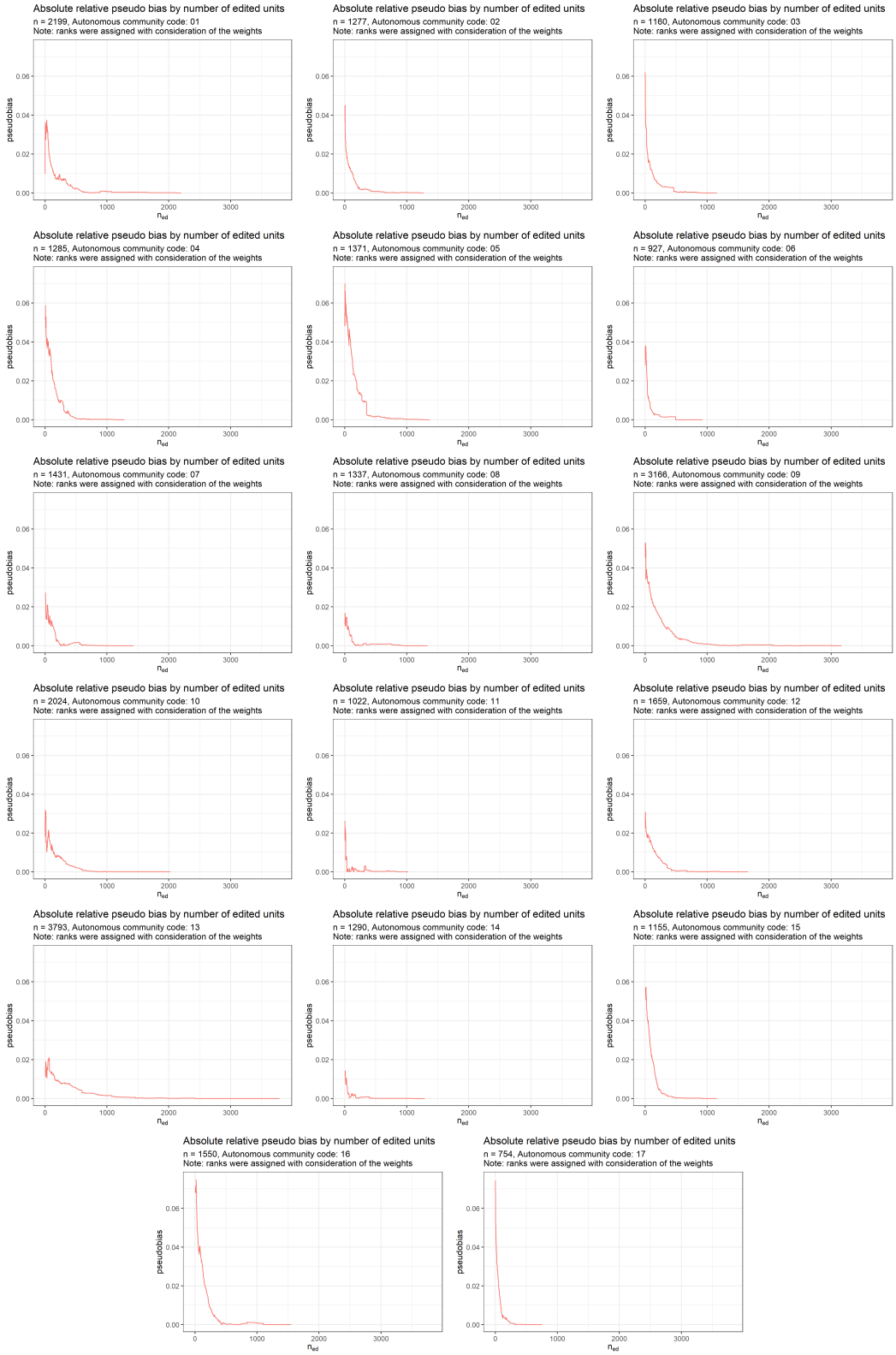

FIGURE A.13: Pseudo-bias by number of edited units by autonomous community

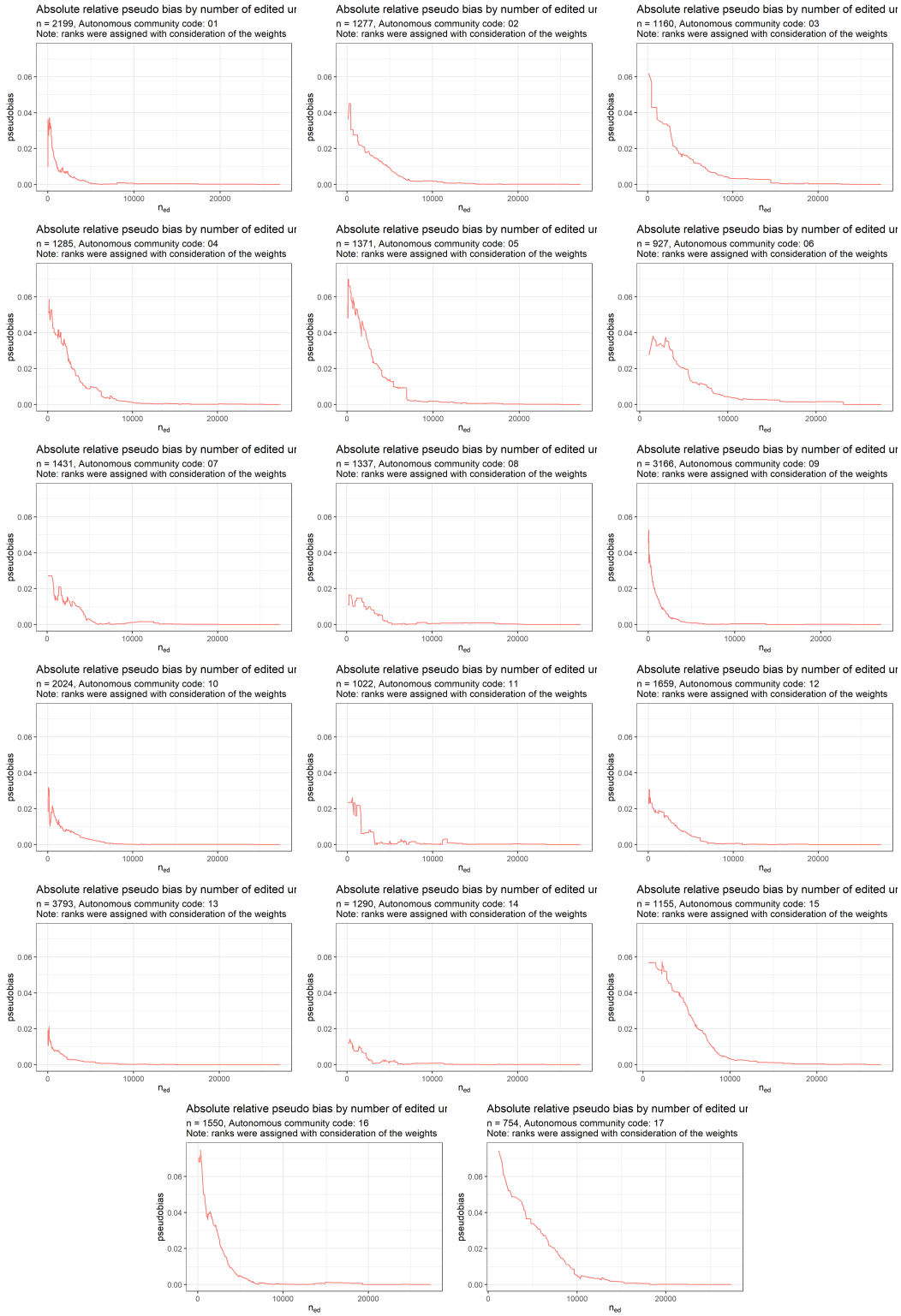

FIGURE A.14: Pseudo-bias by number of edited units by autonomous community respecting the global rank distributions

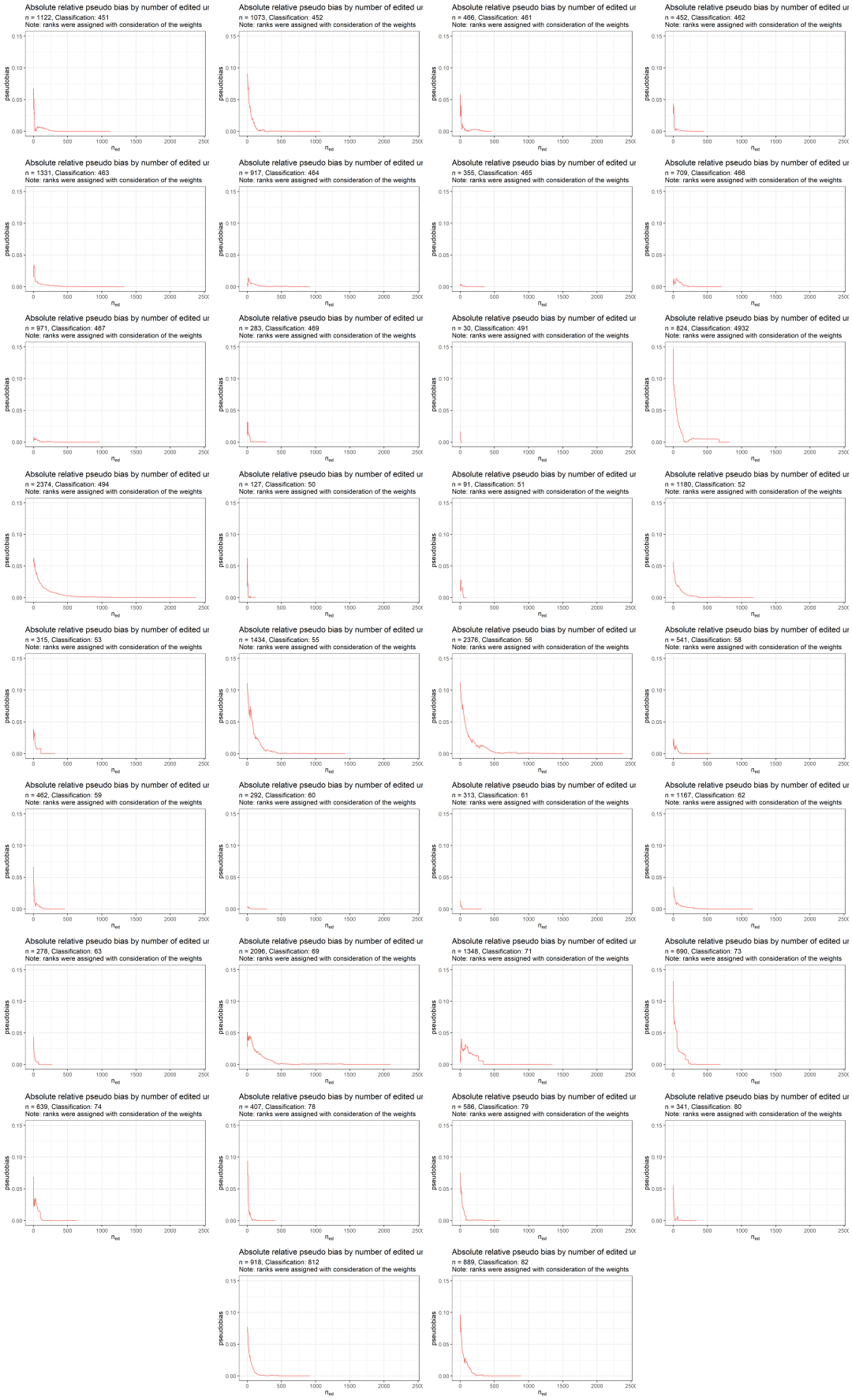

FIGURE A.15: Pseudo-bias by number of edited units by economic activity classification

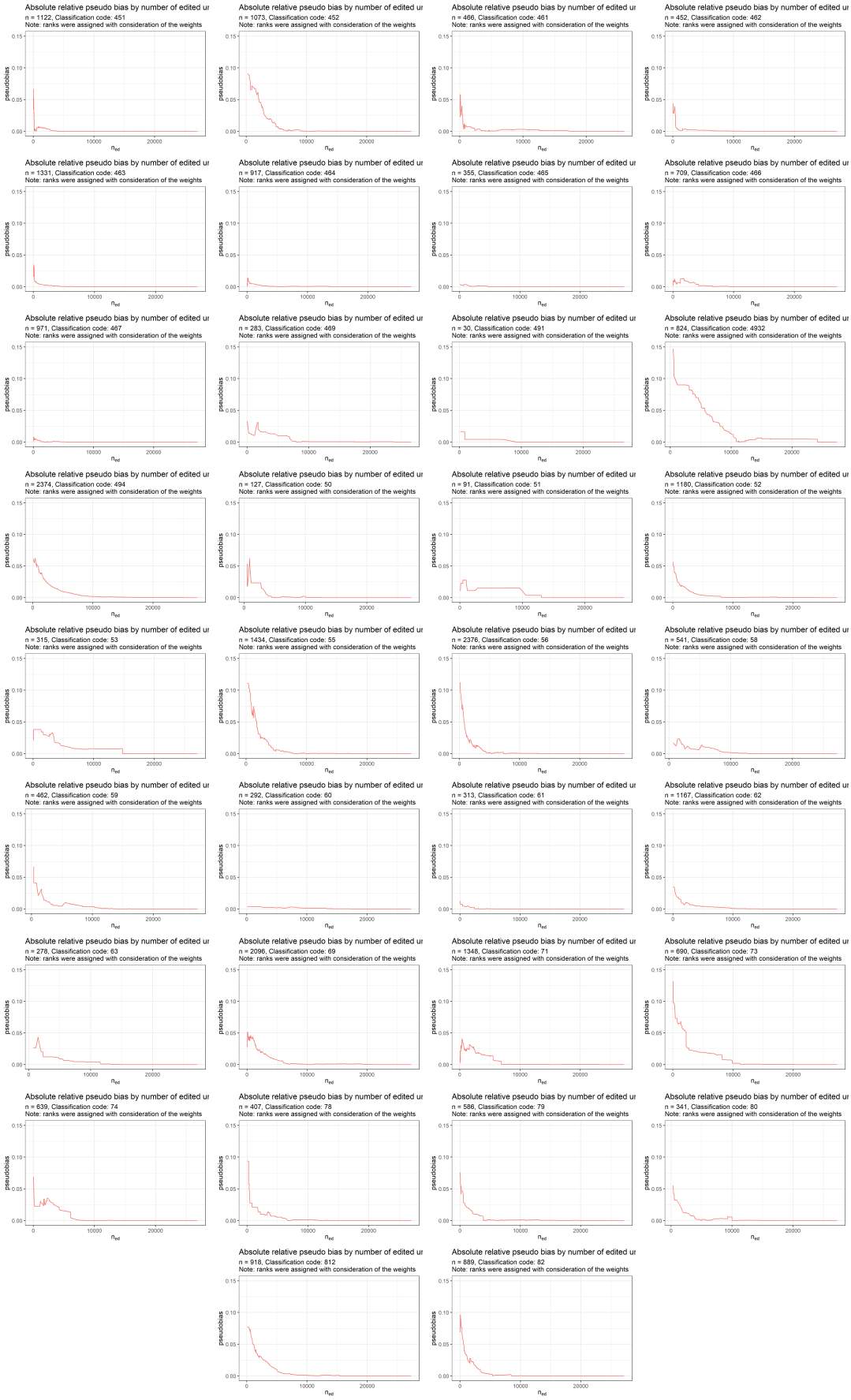

FIGURE A.16: Pseudo-bias by number of edited units by economic activity classification respecting the global rank distributions