POLITECNICO DI TORINO

Master's Degree Course in Management Engineering

Master's Degree Thesis

Measuring Nepotism and Overrepresentation in the Chilean Public Service: Analysis of Paucity and Diversity of Surnames

Supervisor:
prof. Stefano Sacchi

Candidate:
Nicolás Andrés Vega M.
Student Id: 275866

Academic Year 2020-2021

Abstract

This quantitative study assesses the paucity and diversity of surnames of the population in the public service through statistical techniques and tools. Based on the transparency data from the Chilean Civil Service Department and Civil Register Department, this research aims to explore the risk of nepotism and elite capture of several institutions and local ranks in the public service, and in particular, we focus on those groups who were through a merit-based selection. Usability and transparency issues when dealing with public data are approached using smart algorithms and parallel computing tools, we hope that by addressing these problems we open doors to opportunities for new entrants who want to navigate the public data available.

Nepotism and elite capture are detrimental for organizations, especially in the public service. In this research we will show which institutions and local ranks have risk of nepotism or elite capture based on shared last names among public employees. Out of 577 institutions, 355 display a significant risk of either nepotism or elite capture, meanwhile, 87 out of 148 local ranks display significant risk of either nepotism or elite capture. In the same way, we found that probabilities of nepotism risk and elite capture risk in institutions and local ranks depend on predictors such as the number of employees, region (north-south trend), public ranking and wage. Using this approach policymakers can focus their efforts and resources on the most problematic geographical areas and clusters in order to promote fair practices in the public service.

To my family,
my mother, my father,
and the two sisters
life gave me.

Acknowledgements

First, I would like to thank Stefano Sacchi, professor and supervisor at Politecnico di Torino and Naim Bro, postdoctoral researcher at the Millennium Institute for Foundational Research on Data in Chile. Both are excellent professionals who supported me throughout this study, always willing to discuss ideas and show me new ways of thinking. Also, I would like to thank Pontificia Universidad Catolica and all those workers in the Undergraduate Office who made this possible.

To my close friends and all those who always believed and trusted me. This was an amazing chapter of my life, and of course, bigger challenges are coming. Finally, to my mother, my father, my sisters Stephanie and Nathalie. Without you, this would not be possible.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

Governments have to hire workers in order to produce and deliver public goods and services. Moreover, forces that drive incentives inside public institutions are not the same as private ones, then some detrimental practices might emerge especially in less developed countries [1]. We measured in our study two of them: nepotism and elite capture.

Previous research has shown negative effects when observing nepotism and elite capture. Nonetheless, international experiences and prestigious organizations have proved that public service policy and systems based on values associated with fair representation such as meritocracy and diversity can mitigate these both harmful situations, while bringing huge benefits in terms of public service quality and delivery.

In this research, the risk of nepotism and elite capture are social structures that we want to estimate using statistical analysis of public data through a shared surnames approach. For this purpose we used public data about Chilean surnames and public employees in the public service, which due to its nature raises very complex problems in order to be ready to process.

These complex problems forced us to use several statistical and computational tools in order to get the data that is actually useful for our study. In particular, we used well-known libraries in the data science field throughout our research: Jupyter Notebook, pandas, numpy, Dask, plotly and statsmodels. In the same manner, we took advantage of statistical techniques such as Monte Carlo models and logistic regressions which play a key role in discovering insights and evaluating macro and micro situations with huge volumes of data.

Also, cluster segmentation of public employees is made by ranks and public institutions in order to identify the location or composition of risky (in terms of nepotism and elite capture) clusters. Finally, we took a director cluster which was hired through a merit-based system called "Sistema de Alta Dirección Pública" or SADP. Indeed, we analyzed this particular merit-based cluster in order to determine if this cluster that professes the values of merit, equal participation and nondiscrimination generates a different type of surname composition, and therefore, a different risk of nepotism and elite capture compared to non-merit director clusters.

The objective of this work is to show how statistical analysis and smart algorithms can identify areas and clusters with high risk of nepotism and elite capture. Even more, we will estimate which parameters and variables affect both risks and the correlation between these variables.

The output of this research, which is the identification of conflict clusters and their correlation is the first step in a large chain to eradicate prejudicial practices of nepotism and elite capture from the public service. Nonetheless, through the use of these statistical models and computational tools, policy-makers can take the results to focus on specific institutions, ranks or geographic areas, and therefore, build data-driven action plans and effective public policy.

2 Literature review

2.1 Jump the queue through nepotism

Appointment of family members in politics, business and public positions is very frowned upon in some places, but is not uncommon to observe. In April 2018, after Sebastian Piñera assumed for the second time as the president of Chile, he appointed his brother Pablo as ambassador of Argentina, a serious and critical charge in Chilean diplomacy [2]. In the same way, President Donald Trump appointed his daughter and his son-in-law as White House's advisors during his administration in 2017.

This kind of appointment, in which a person or group appoints one relative to a certain position in an organization is generally called nepotism. Even though nepotism may bring some benefits to organizations [3], it is often labeled as unfair, unethical and unprofessional [4] depending on its context.

There are several types of nepotism, and some of them are even perceived as acceptable [5]. For instance, royal families, who not a time ago were quite popular, are the most clear example of nepotism structure where high power positions are passed from one generation to the other.

Family businesses are the same, it is not uncommon to see organizations in which the CEO or any other C-level executive is a close relative of the owner or board of directors. In fact, 500 Fortune's companies analysis made by Zajac and Westphanl showed that a board of directors is more likely to select a CEO who was considered as an in-group member [6], which is not surprising, since members of a certain group tend to act in a way that brings benefits only to their own circle [7].

Same situation of acceptable nepotism happens for family holdings or family offices, which may include a diversified portfolio of organizations or assets that are almost exclusively managed by family members. In Chile, we have wide examples such as: Solari's family office "Corso", President Piñera's family office "Bancard", Matte's family office "Porto Seguro" among other gigantic ones [8].

One may think that nepotism becomes a tricky concept to label as bad or good when we base our judgement on these diverse situations which we had described before. Nonetheless, our goal is not to question if these situations described above are fair or not, but rather if nepotism is valid when you are dealing with a public organization, and even more specifically, the public service.

Favoritism of a relative in a hiring process may be considered an act of nepotism [9]. Indeed, for most scholars just the fact of hiring relatives of current employees of

an organization might be labeled as nepotism, without taking into account the relative's qualifications for the position [10]. Then, nepotism in a more narrow sense should be: hiring someone's relative who has not enough qualifications for the job that he is applying for [11]. But even though some situations can be labeled as nepotism, we might focus on the people's perceptions of it.

Perception of nepotism can influence the organization's ability to hire and attract a highly qualified workforce [12]. In the same line, it has been associated with poor organization management [13], lower worker effort, lower firm performance [14], and poor bureaucratic performance [15]. Organizations known as nepotistic ones may be considered not attractive to apply for, specifically for those who highly value merit principles and fairness.

And for those who are already working inside the organization, perceived nepotism may increase counterproductive behaviors such as decreased morale [16], bad-mouthing the organization and manifestations of intentions to quit [17]. Moreover, organizations that are perceived as unfair are associated with lower organizational commitment and job satisfaction [18], which is clearly the case of the Chilean public service at least in these last two indicators (see section 2.6).

Researchers who have studied the perception of nepotism in organizations, have found that a higher proportion of genetic overlap among people in the same organization was associated with a higher perception of nepotism [19].

Then, our proxy to measure the perception of nepotism is by identifying public servants that work in the same public organization and share surnames. This strategy allows us to label every single public organization or rank cluster of workers as homogeneous or heterogeneous by comparing them with a fair representation of local surnames multiple times.

The use of sharing surnames approach is not brand new. Research made in 2010 for the municipal elections in the Philippines showed that individuals who share one or more family surnames with local elected officials were more likely to be hired in better occupations, compared to individuals who shared surnames with losing candidates [20]

In the same way, Durante et al. (2011) found a higher concentration of surnames in Italian universities compared to the Italian population (which will be the approach taken by this study), and this concentration increased in low civic capital regions [21]. Finally, part of our study intends to capture those clusters who present a significant perception of nepotism using the shared surname approach.

Although nepotism may happen frequently and their consequences can be detrimental, there are some well-known cases such as Singapore, which illustrates very well how anti-nepotism policy, but also robust meritocratic policy may bring extensive benefits [22].

2.2 About meritocracy

The panacea to solve the latest problems and side effects showed in the context of the Chilean social protests of October 2020 and the pandemic is far away. Nonetheless, it seems that when citizens talk about the need for structural reforms and changes in the Chilean public system, they reach a consensus.

Inequality gaps, few access to opportunities, and lately, the government failures have generated a huge debate which have led to a strong concept: meritocracy, and although Chileans perceive that the society lacks meritocratic features, they still consider meritocracy a suitable system for wealth distribution [23]. In the same way, the lack of connection by important public authorities and managers has left footprints of a big necessity of representation, most especially for those who are more vulnerable in the current situation of health and economic crisis.

Meritocracy involves not only merit and effort, but other variables which are intrinsic to the society where citizens live. In particular, meritocracy is not really a real version of itself if it considers non-meritocratic aspects at the moment of shaping an organization that gives service to the whole community, such as the public system. Even more specifically, a true meritocracy has to assure the assessment of a person's merit without taking in consideration aspects such as: ethnic origin, disability, age, gender, sexual orientation, religion or belief.

Under this last statement, an organization has to be representative of the population which it serves, not only to pursue a better financial performance or productivity, but also to establish important ethical and non-discriminatory values [24].

In its simplest way, a meritocratic system is defined as a social order in which rewards are exclusively based on the individual excellence under the idea of merit. And merit "is the sum of intelligence plus personal effort" [25]. This last definition was made by Michael Young in 1958, and it may feel a little incomplete compared to the way that citizens understand meritocracy today.

In the context of public service, meritocracy could be defined as a system where every position and prestige is fairly distributed among those who possess the best suitable skills or merit (studies, professional career, performance among other variables) for the job.

Nonetheless, a system in which the worker with the most suitable skill set gets the job is not completely fair. According to the Yale's researcher, Daniel Markovits, there is a "trap" in meritocracy that has been studied in countries like the United States, where inequality in education's access provoked an unequal workforce. Then, the incomes of the new meritocrats enabled them to buy an unequal education for their children, which generated the next unequal workforce generation and so forth, and this phenomenon has led to an increase of inequality gaps in each generation [26]. In other words, is a hereditary meritocracy in which the elite's children, in the context of meritocratic systems, are more likely to over perform under meritocratic standards in employment and education parameters [24].

Then, it is necessary to bring an additional concept to the big picture which suggests that everyone should play in the same field and under the same rules: equality of opportunities.

To achieve the meritocracy that everyone wishes, we need: quality in education, health access and a good public transportation. All of them contribute to provide citizens with equal opportunity for advancement. And in the same way, we need transparency, accountability and regulation, meritocracy does not exist in isolation [24].

In effect, now the concept of meritocracy is better defined, and it is known as "true meritocracy" in the literature: a system in which "everyone has an equal chance to advance and obtain rewards based on their individual merits and efforts, regardless of their gender, race, class, or other non-merit factors" [27]. And merit is defined as an alignment or trade off between achievements and performance [28].

Benefits of meritocracy are not unknown, authors have identified that it reduces corruption [29], limits patrimonial networks [30], fosters economic growth [31], increases personnel performance [32], among others. Also, important institutions such as the World Bank have suggested a meritocratic system in the public service to help bring high-quality staff, which can confer prestige on public service positions and motivate a good performance among institutions [33].

At the same time, there is another important dimension which is deeply connected to the way of seeing a true meritocracy, and it must be brought to the table: diversity. The diversity as a representative element of the population can generate several positive effects in any public system.

2.3 Diversity as element of representation

There are a lot of definitions of diversity, and its differences are not only semantic, but also reflect the strategic priorities of governments. Even so, definitions of diversity

can be summarized in three main groups:

- Diversity as equal opportunities: it is seen as prevention of discriminatory practices in terms of gender, age, ethnicity, religion, belief, sexual orientation, political opinion, disability, physical appearance or other. Therefore, it guarantees neutrality of the human resource management, allocating value to people by their own merit regardless of the aspects mentioned before.
- Diversity as an asset, it proposes to see, understand and appreciate the benefits that can bring different experiences of life, competencies, socioeconomic and cultural backgrounds to the performance of the public service.
- Diversity as social inclusion, which alludes to the long-term strategic work to ensure structural changes in organizations. In particular, it intends to make use of every person's relevant competence for the organization.

Most of the OECD governments have adopted one or more of these three groups with the intention of pursuing a better representation of the society, and this is precisely which is intended in this research: be the first to measure and assess representation (through shared-surnames) in the Chilean public system using cutting-edge algorithms to process public data.

The first analysis of the benefits which diversity can bring to organizations are seen in the private sector, mainly motivated by increases in performance inside of organizations, and of course, their financial success compared with their pairs in the local industry.

Evidently, most studies are based in completely different contexts compared to the one in which people are living today, where businesses are shut down, flights are cancelled and borders are closed due to coronavirus. In the same line, these studies do not consider new global issues related to the new wave of demonstrations against systemic racism. Without any doubt, it is a time of uncertainty, but also a time for reflection that hopefully restructures individuals, communities, organizations and governments compositions to manage new necessities which will emerge in the so-called "new reality".

In particular, studies have focused on the relationship between diversity (defined as gender and mixed compositions related with different ethnic origins and races) and financial success of companies worldwide. Their results have shown a significant statistical correlation between a diverse leadership with a better financial performance: companies in the first quarter of racial/ethnic diversity were 35% more likely to get higher financial returns compared to the median of the national industry [34].

In the same line, researches are found in Latin America about diversity, its effects in organizations, and how this way of representation may affect key financial indicators and organizational health.

More precisely, in companies that foster diversity employees tend to experiment more freedom in both their identity and their way of working. In fact, when employees perceive that their companies are committed to diversity it is 111% more likely to report that they can be themselves while they work, which probably encourages and empowers them to participate and contribute [35].

Also, these same employees are: 152% more likely to report that they can propose new ideas and new ways of doing things at work, 76% more likely to say that their organization uses feedback from customers to give a better service, 77% more likely to agree that their organization applies external ideas to improve their performance, 72% more likely to report that their organization consistently improves their way of doing things and 64% more likely to say that they can collaborate sharing their ideas and better practices [35].

On the other hand, the levels of a company commitment with diversity are strongly correlated with the capacity of its leaders to cultivate trust and high-performance teams. In financial aspects the story is the same: Latin American companies perceived as diverse in terms of gender are 93% more likely to over perform their pairs in the industry [35].

2.4 Diversity in public service

One of the main premises of this research is that a diverse workforce contributes to make a strong and fair public service which understands and fulfills better the expectations of its citizens. Thus, the research hopes for a future Chilean public system with a workforce characterized by social mobility, diverse personal backgrounds, experiences and competencies in order to satisfy social necessities. Diversity is not just an end in itself, but a mechanism to reach wide and inclusive public policies. Therefore, social situations such as nepotism and overrepresentation attempt against our hopes.

When the goal is to improve the representation of several o certain social groups inside the government, diversity plays an important role in maintaining key public values, increasing management efficiency, improving effectiveness of public policies, fostering social mobility and, strongly, ensuring quality in public service's delivery [36].

For most of OECD countries, diversity in public service has become a top priority since it helps to reach political and social objectives, for instance: social mobility, equity and better service delivery. There seems to be a consensus in these countries in pursuing

diversity as a strategy for the preservation of central values of public service, such as justice, impartiality, and particularly, representativeness.

Also, diversity may help to advance reform agendas and promote good governance practices in order to improve relations between governments and their citizens, generating trust networks [36], which could be needed since social protests are more present than ever before.

The action plan to reach a more representative workforce is deeply related to hiring processes that are more fair, transparent and flexible. Therefore, it may attract talents from a variety of backgrounds, experiences and perspectives. These mechanisms of selection and recruitment is when the key composition of the workforce in the public service is at stake.

Specifically, better recruitment processes are related to mechanisms that may: i) diversify communication channels to reach a wider audience; ii) motivate citizens to apply to the public service offers; iii) lighten selection processes and criterion to make it more inclusive, but still focusing in analytic and quantitative skills and required competencies to the job; iv) facilitate the integration and retention of the new public service workforce [36].

2.5 Elite and public service

Diversity seems to bring huge benefits in organizations, including public service. One may think that in a sample with highly diverse surnames there should be more diverse people since surnames are associated with social information about ancestry, but this is not always the case.

Recent research made in Chile shows that social status can be a very important factor that shapes the Chilean society. In fact, surnames which are associated with higher incomes, tend to group in ethnic clusters, and even more, the study concludes that there is a positive correlation between the socioeconomic status and the diversity of surnames, which means that "the higher the income, the more diverse the surname composition" [37].

Let us take the case of public service again, as we described above, different views, experiences and perceptions may bring benefits to the public service. But it can happen that groups in the public service with high decision-making power are predominantly very diverse in terms of surnames, which may reflect elite clusters.

High representation of elite public workers makes a suitable context to facilitate elite capture which can produce very detrimental results. Elite capture's issue is observed in

most developing countries, in which a subgroup of the community who possess greater power and information over the decision-making process, in this case high public positions, lead to a civil society failure, this means, that self-interest prevails instead of community interests [38].

We can find the roots of the belief that local elites steal and capture resources going back to the Federalist papers. In the same way, this belief has been explored in recent years by several scholars in order to identify its causes and effects on different fields.

The long history of elite capture extends its presence across a diverse range of areas such as corporate networks [39], politics [40], globalization [41], public land distribution [42], information distortion [43] and caption of foreign aids [44].

Elite capture may result in a situation where certain groups of the community, usually the poorest ones, reduce their access to public goods and services, indeed, previous research concluded that elite capture reduces social welfare [45].

In the same way, its effects are commonly related to bias in public resource allocation so as long as elite capture is present in the public service, the welfare impact will not be Pareto Optimal [46]. Moreover, studies demonstrated that elite capture can lead to economic marginalization of poor individuals, and foster existing levels of poverty and inequality [47,48].

Therefore, it seems necessary to first: identify the potential public cluster (i.e. institutions or ranks) in certain areas of the public service that can show a high likelihood or risk of elite capture. The identification is essential to propose public policy to minimize elite capture, which may allow to achieve a more equitable and sustainable governance outcomes [49].

As well as nepotism, elite capture will be measured by measuring the diversity, or what we called "overrepresentation", through the shared surnames approach which seems suitable for the Chilean situation due to the surname cluster composition of high incomes groups [37].

2.6 The Chilean case

There is no consensus about how a meritocratic or diverse bureaucracy has to be measured. Nevertheless, it is clear that following a model under the ideas of meritocracy and diversity may bring huge benefits to the public service, and in particular, to the delivery of public goods and services to the Chilean population. This study aims to

propose and foster meritocracy and diversity, but also, target risky clusters in order to avoid nepotism and elite capture.

Bureaucratic structures and their configuration can be explained and analyzed through two mechanisms: the hiring and promotion process, which ideally should be based on meritocratic criterion and other typical aspects of a meritocratic organization.

In particular, the Chilean case is very interesting taking into account that 61% of the employees of the Chilean public system did not go through any written test, psychological evaluation or selection committee, and although this percentage goes down to 50% when it takes only those with 0 to 5 years of service, it is still half of them [50].

The recent survey made by the Chilean civil service department indicates that those selected through more objective evaluation processes than their peers (written test, psychological evaluation or selection committee) have better integrity indicators, are more satisfied with their job and are more willing to collaborate [50].

Recommendations to solve this lack of selection standards point to prioritize consistently more meritocratic criterion (not political nor personal criterion) in the decision-making at the moment of hiring, promotion or firing personnel, including universal application for public contests and offers.

In fact, public contests have to be a universal rule at the moment of applying for a public service role because only 36% of public employees hear the offer from a public contest, and data has shown that civil servants who were recruited through a public channel tend to demonstrate higher levels of ethics, integrity and commitment. On the contrary, employees with personal or political connections at the moment of hiring tend to behave in a significant less ethical way [50].

The data presented above shows that public service's standards for hiring in Chile are not enough. Because of that, this research pretends also to focus efforts on merit-based selection groups, in particular, the "Sistema de Alta Dirección Pública" (SADP). SADP was implemented in 2003 to hire and select top public managers in certain government areas, moreover, SADP has as main goal to hire suitable public managers based on qualifications, effectiveness and efficiency, following the principles of: merit, equal participation, confidentiality, non-discrimination and transparency [51].

One of the ideas behind the research is to understand how one of the few merit-based systems is affecting diversity, and more precisely, representativeness in top positions through its surname composition, and how this particular "merit" cluster can be compared with other clusters inside the public service.

The analysis will be conducted using public data sets available in the Civil Service Department of the Chilean Government. More specifically, the data set contains information about workers hired through different contracts and their information about names, department, wage, role, profession among other valuable variables. In the same way, the research uses data sets of Chilean last name's frequencies by municipalities provided by Millennium Institute for Foundational Research on Data.

3 Motivation

Studies about co-occurrence of surnames is not necessarily a brand-new method, but it has brought interesting findings. This approach has been used in the academic field, and more specifically in academia. Durante et al. analyzed Italian scholars through their frequency of surnames, clustering by geographic area. His results showed that the homogeneity degree in Italian academia is much higher than the expected in a random situation, in particular, for some areas and institutions [21].

In the same way, our research has followed a very similar approach to the one used by Stefano Allesina, an italian researcher of the Department of Ecology and Evolution at the University of Chicago. Allesina [52] shows how diverse disciplines in Italian academia with high likelihood of nepotism can be detected with statistical techniques and the paucity of surnames. Additionally, these techniques are simple to use with the current stack of technological tools available.

Nonetheless, there are differences between the Chilean and Italian to bring to the table before using the same kind of approach. In a broad sense, under Chilean law when a person is born, then she or he is registered with two names and two surnames. The names deliberately can be chosen under certain constraints and the surnames are registered as the paternal surname called "first" surname, and maternal surname called "second" surname.

For example, if the paternal surname of Pablo's father is Reyes and the paternal surname of Pablo's mother is Rojas, and they decided to call his child Pablo Antonio, then his child's full name will be Pablo Antonio Reyes Rojas. Additionally, in the Chilean case women maintain their both surnames, which do not occur in other countries. So under these last statements, the detection of possible nepotism can be only associated with a relationship between father-child, sibling-sibling and between paternal relatives.

Another difference between the Chilean and Italian context is noticed when we are analyzing the surname's frequency. The difference is explained by the distribution of surnames, where it is more frequent to find repetitions of surnames in the Chilean population, instead of the Italian population.

The most frequent surname in Chile is "González" and it represents around 2% of the Chilean population, while the most frequent Italian surname is "Rossi" which represents around 0.12% of the Italian population. Then, if we analyze carefully how the distribution of most Italian and Chilean surnames is, one could expect more repetition of surnames in Chilean samples rather than Italian samples, being those with equal numbers of people.

It is important to establish that sharing surnames does not necessarily imply a familiar relationship, but it can be used as a proxy for nepotism in cases where paucity of surnames is significantly lower compared to a representative sample, i.e, random in this case. In the same way, backed by previous research overrepresentation is a proxy for elite capture.

These differences that we have mentioned before does not mean that we cannot take the approach used by Allesina to detect nepotism in the Chilean public service. In fact, we can use the same technique and procedure to explore a new dimension: the overrepresentation, or high diversity of surnames (opposite to the paucity of surnames).

As we described, the Chilean population has a lot of frequent surnames, therefore, is very rare if a certain institution or rank has a lot of distinct surnames, i.e. diverse surnames or overrepresentation. Recently, Bro and Mendoza et al. [37] suggested that these diverse surnames groups tend to be more associated with Chilean traditional high classes.

Specifically, the members of these groups tend to have less frequent surnames, possess a high socioeconomic status and be overrepresented in politics [37]. Thus, this potential rare source of high diversity, which is a concept completely opposite compared to the diversity described in section 2.3 and 2.4, indeed, it allows high risk of elite capture which may have negative effects for the public system, especially for social mobility and the understanding of social needs.

Finally, our intention is to discover who and where these both phenomena may happen in order to prevent certain clusters to control resources from the public system, avoiding capture state by private interests [53] instead of social interests.

4 Data

The main databases and sources of information for the study come from:

- Data about Chilean surnames and its frequency by Municipality (D1). Gathered by The Millennium Institute Foundational Research on Data.
- Data about public service employees under different types of contract (D2). Download and publicly available in the Transparency Portal (Portal de Transparencia) of the Chilean Government.
- Data about active employees in SADP (D3). Download from the web page of the Civil Service (Servicio Civil) of the Chilean Government.

From this point, we will call i), ii), iii) as D1, D2 and D3 respectively.

4.1 Data cleaning

Before doing any expensive operation with data, one of the most important steps of data analysis is the data cleaning.

For this case, when we did a light analysis of our sources of information, and consequently we found several ways and reasons to preprocess our databases (D1, D2 and D3) in order to clean them:

- Duplicates: duplicates are found in all three gathered databases, therefore, we proceed to remove duplicates, because we are analyzing a specific moment of time, therefore, there is no way to find two equal rows (equal attributes in both rows) in the same database. For instance, in the case of databases which contain public employees, we remove duplicates because it is not possible (by law) to be working at the same time twice or more times in one or more public institutions.
- Uppercase and lowercase: both are human mistakes. It is very often to find a mistake when a person is trying to input a row in a database. In the same way, sometimes there is no agreed format to upload data. Both can be mistakes or lack of standards which can cause inconsistency in the database at the moment to match words (see section 4.3.3). In the case of surnames, and other qualitative variables which contain letters are transformed to uppercase for further analysis. This step is essential at the moment to do a matching between two people or grouping by surname.

- Tildes: tildes also are another element which is very frequent in typing error of qualitative variables, and for the purpose of this research, we will remove all tildes from the set of surnames and its related qualitative variables. Then, this implies that we are assuming that a surname is equal to the same surname with one or more tildes in any of its letters.

- Incomplete data: there is data that does not have all its information in certain attributes. Therefore, numeric variables of an incomplete row are replaced by the median in the respective attribute. In contrast, if the variable is categorical or qualitative, then the missed value is replaced by the mode of that variable. Incomplete rows are very few, then the impact in the data is very low in this case. Finally, this replacement helps us to maintain the consistency of the data for further analysis.

- Outliers: from the lightweight analysis we determined outliers in the data which can be explained mainly by typing errors of certain attributes of the database. The most notorious outliers were the ones related with wages, being those extremely high for the occupation or charge of the person. Those cases are impossible to be real, because Chilean law establishes a certain range of wages by public grade or rank. Then, to solve these outlier values, they are replaced by the median of the outlier variable using the most suitable cluster of the dataset. For instance, if we found that a certain driver in a public hospital is earning €70.000 monthly, and the median wage of drivers in public hospitals is €700 monthly. Then, we identify the row as outlier, and the value €70.000 is replaced by €700.

4.2 Data about Chilean surnames and its quantity by municipality (D1)

Data about surnames and its distribution in the whole national territory were provided by the Millennium Institute for Foundational Research on Data. At the same time, the Institute sent a request to the request channel of the Transparency Portal of the Chilean Government, which provided the data from the Civil Register. This data contains all surnames in the country by municipality in the 2018.

The data contains all the Chilean surnames and its respective frequency by municipality (346 municipalities) and region (16 regions in the national territory). After data cleaning, we manipulate D1, in this case, we grouped by surname (see Figure 1 and 2). This data is available for either paternal or maternal surnames in separated datasets.

In this way, the number of total rows in the data are 738.350 with a unique set of 95.619 paternal surnames. In the case of maternal surnames, the number of total rows

is 719.703 with a unique set of 103.531 maternal surnames, which determines a slightly bigger heterogeneity of surnames in maternal surnames.

REG	COMUNA	AP PATERNO	CANTIDAD
1.00	ALTO HOSPICIO	MAMANI	66.00
1.00	ALTO HOSPICIO	CASTRO	15.00
1.00	ALTO HOSPICIO	CHALLAPA	13.00
1.00	ALTO HOSPICIO	CHOQUE	13.00
1.00	ALTO HOSPICIO	FLORES	13.00
...	...	...	...
16.00	SAN CARLOS	ZAMBRANO	1.00
16.00	SAN CARLOS	ZAPATA	1.00
16.00	SAN CARLOS	ZURITA	1.00
16.00	SAN IGNACIO	CHAVEZ	1.00
16.00	YUNGAY	GOMEZ	1.00

738350 rows × 3 columns

Figure 1: Information gathered from database D1. Its attributes are the following: paternal surnames (AP PATERNO), its respective frequency (CANTIDAD), municipality (COMUNA) and associated region number (REG).

REG	COMUNA	AP MATERNO	CANTIDAD
1	ALTO HOSPICIO	MAMANI	53
1	ALTO HOSPICIO	CASTRO	16
1	ALTO HOSPICIO	CHOQUE	14
1	ALTO HOSPICIO	CHINO	11
1	ALTO HOSPICIO	GARCÍA	10
...	...	...	...
16	SAN CARLOS	ZAPATA	1
16	SAN CARLOS	ZÚÑIGA	1
16	SAN CARLOS	ZURITA	1
16	SAN IGNACIO	RIQUELME	1
16	YUNGAY	URIBE	1

719703 rows × 3 columns

Figure 2: Information gathered from database D1. Its attributes are the following: maternal surnames (AP MATERNO), its respective frequency (CANTIDAD), municipality (COMUNA) and associated region number (REG).

Later, the next operation that we have to apply to the data is to group paternal and maternal surnames by region, which means, aggregate surnames of several municipalities of a certain region, and add their frequencies. A region is the bigger unit in geographic terms to characterize the population based on their surnames, and it will be the one used for this research (see Figure 3).

Out[18]:

REG	AP PATERNO	CANTIDAD
13.00	GONZALEZ	172857
	MUNOZ	119926
	DIAZ	90809
	ROJAS	86148
	PEREZ	71250
...	...	...
5.00	BURATOVIC	2
10.00	FELLAY	2
13.00	POMEROY	2
10.00	ENDRUSSAT	2
	FONCHZIK	2

148415 rows × 1 columns

Figure 3: Information gathered from database D1. Its attributes are the following: Paternal surnames (AP PATERNO) and its frequencies (CANTIDAD) aggregated by region (REG).

This aggregation by region will help us in the following steps of the research to randomly select samples from one or more regions in order to simulate people for a certain public institution or rank.

The information of this database is a key differentiator, because it makes the research differ from previous research which used a similar approach. In particular, most previous research did not have available at that time the information about the national distribution of surnames by geographic area, therefore, they created samples using the pool of data which they are trying to analyze. For example, in Allesina's work, he used the total pool of scholars to take samples [52]. In our case, we use the total population of the country to get those samples, thus, it opens a possibility to make an analysis deeper and more accurate when we are trying to simulate fair representation.

4.3 Data about public service employees under different types of contract (D2)

4.3.1 Size of databases

These four databases (i.e. four files by each type of contract) have a particularly huge size which makes it difficult to process in a personal computer using traditional methods (see Table 1).

Database file	Size
Indefinite contracts	2.37 GB
Fixed-term contracts	3.64 GB
Fee contracts	2.01 GB
Other special contracts	1.61 GB

Table 1: Public data available gathered from the Transparency Portal of the Chilean Government. Files ranges between 5 and 11 million rows.

The size makes infeasible the intended manipulation of data through classical analysis methods using a common personal computer. Therefore, we have found a big obstacle to process the public data available, which may explain the fact that nowadays there are few people or institutions taking this data to process and visualize.

One of the institutions that we could contact for this study was "America Transparente", a foundation that created a tool called "Reguleque" [54], a search engine to navigate through this transparency data in a friendly way. This foundation is aware of our current study, indeed, we have established a very active collaboration since we both are using the same data, and went through the same pains.

As we mentioned before, these databases are huge because they contain historical records from 2013 until now. Logically, the data that we want to use is not all, because we do not want to know how things were, but rather how things are or at least get the most recent records that can characterize the current situation in the public service. Consequently, we will select a particular moment in time: April 2020.

Therefore, the main need is to filter these huge volumes of data to obtain only those of interest for our research, these being the data of a certain month and year. However, the problem does not end there, even trying with one of the most popular libraries of data science, pandas, is not enough since the data cannot be handled in memory (from 5 millions to 11 million rows).

Given this problem, searching for another tool more powerful (than pandas) to process huge volumes of data is strictly required. So we used the library Dask [55] which

completely fulfilled our technical requirements. More specifically, Dask can handle data processing tasks which our RAM memory is not able to handle only by itself. This can be possible because Dask uses multiprocessing in parallel, which means in simple words that it breaks a huge computational task into small ones that are processing in parallel. This particular feature allows us to process up to 100GB+ datasets with no local problem. Moreover, Dask is compatible (and shares a very similar API) with Pandas, which will be the library used for the next steps in the research with our databases already filtered.

As a result of the use of Dask's data filtering, the size of the files which contain the information goes down considerably (see Table 2), easing the processing and manipulation for further steps of the data analysis. These filtered files are the ones that will be used multiple times during the course of this research, remembering that these databases have important metrics such as surnames, institutions, remunerations, ranks, occupations among others.

Database file	Size
Indefinite contracts	56 MB
Fixed-term contracts	104.9 MB
Fee contracts	36.7 MB
Other special contracts	35.8 MB

Table 2: Public data available and filtered using Dask library.

4.3.2 Origin and exploration of data

The information about public service employees of several public institutions of the Chilean Government is available and open for everyone in the webpage of the Transparency Portal (*www.portaltransparencia.cl*) in the section of "Reports and Data" (Datos e Informes), and then more precisely, in the path "Open Data" (Datos Abiertos).

Specifically, the available data in this portal that we need for the study is the information about: employees hired with indefinite contract, fixed-term contract, fee contract and through specific work laws (all of them in separated files). As we have mentioned before, these files contain historical data from 2013 until now of the whole current and old public personnel month by month (see Figure 4).

	camino	organismo_nombre	organismo_codigo	fecha_publicacion	anyo	Mes	Tipo Estamento	nombre	grado_eus	tipo_calificacio
13532	/Municipalidad de Llanquihue/2020/Mes de Febrero	Municipalidad de Llanquihue	MU142	2020/03/13	2020	Febrero	Directivo	VERA URIBE MARINA LOURDES	7	ASISTENTE SOCI
13533	/Municipalidad de Llanquihue/2020/Mes de Febrero	Municipalidad de Llanquihue	MU142	2020/03/13	2020	Febrero	Técnico	ALVARADO MANSILLA IRENE SOLEDAD	11	TECN.ADMINISTRATI
13534	/Municipalidad de Llanquihue/2020/Mes de Febrero	Municipalidad de Llanquihue	MU142	2020/03/13	2020	Febrero	Administrativo	ALVARADO TALMAR JUAN CARLOS	14	CHOFE
13535	/Municipalidad de Llanquihue/2020/Mes de Febrero	Municipalidad de Llanquihue	MU142	2020/03/13	2020	Febrero	Alcalde	ANGULO MUÑOZ VICTOR RUBEN	5	CONTADOR GENER
13536	/Municipalidad de Llanquihue/2020/Mes de Febrero	Municipalidad de Llanquihue	MU142	2020/03/13	2020	Febrero	Directivo	ANTIMAN HUERQUE MARCIA ELIZABETH	7	ADMINISTRADO PUBLIC

Figure 4: Sample of the available data in the Transparency Portal about indefinite contract employees in the Chilean public service.

The attributes which are shared in these fours files (fixed-term, fee, indefinite, work law contract) that contain the data about public workers are the following:

- *camino*: is the specific route in which it can be found that row in the webpage of the Transparency Portal of the Chilean Government. It does not have a consistent format.
- *organismo_nombre*: name of the public institution where the person is a public worker.
- *organismo_codigo*: internal code of the public institution where the person is a public worker.
- $fecha\_publicacion$: date when the row is published in the records.
- *anyo*: makes reference to the year when the worker was working in the public service.
- Mes: makes reference to the month when the worker was working in the public service.
- *Tipo Estamento*: rank of the worker according to the public hierarchy and the occupation of the worker.
- *nombre*: full name of the public worker, this means, first name, second name, paternal surname and maternal surname. It follows the format "paternal surname maternal surname first name second name".
- *grado_eus*: is the degree (number) which reflects the remuneration scales determined by the Chilean law.

- *tipo_calificacionp*: type of qualification of the public worker, in general makes reference to the profession or educational level of the public worker.
- *Tipo cargo*: is the name of the charge or role of the public worker.
- *region*: region where the institution (the one that the public worker belongs) declares its location.
- *asignaciones*: monetary incomes or additional benefits (without taking into consideration remuneration) which belong to the public worker.
- *Tipo Unidad monetaria*: currency of column "asignaciones".
- *remuneracionbruta_mensual*: net remuneration of the public worker.
- *remuliquida_mensual*: gross remuneration of the public worker.
- *diurnas*: extra hours worked during the daytime.
- *nocturas*: extra hours worked during the nighttime.
- *festivas*: extra hours worked during the holidays.
- *fecha_ingreso*: employee's registration date
- *fecha_termino*: employee's termination date
- *observaciones*: observations added by the institutions at the moment when the public worker is hired.
- *enlace*: empty field without references.
- *viaticos*: travel and other related expenses received by the employee.

Obviously, not all variables will be useful for our main study purpose. Nevertheless, for the previous analysis most of these variables will be very crucial to clean, transform (if needed) and most importantly understand the data we are using, and in particular, be smart to select those future variables for our statistical model.

More specifically, we did an exploratory analysis of the variables to identify the most interesting insights at first glance. Then, after doing the exploratory analysis of the data, we figured out that we have to focus our attention on these variables:

- *nombre*: to obtain paternal and maternal surnames of public workers.
- *organismo_codigo*: to cluster the public workers by public institutions.

- *región*: to cluster the public workers by region.
- *Tipo estamento*: to cluster groups according to their ranks in the public hierarchy.

Exploratory analysis was an important step for this study because it helped us to understand the structure of the data and its potential variables for the statistical model that it will be used in further sections. In D2, we calculated using the four types of contract the following insights as a way to understand the data:

- Names and role or charge of the n employees with the best remuneration in the public service.
- The n most frequent surnames, its frequency and its proportion in the public service.
- The n most frequent Mapuche (the biggest native population) surnames, its frequency and its proportion in the public service.
- The n most frequent Aymara (second biggest native population) surnames, its frequency and its proportion in the public service.
- The n most frequent surnames of other native populations in the public service, its frequency and its proportion.
- The n most frequent surnames of certain ethnic clusters and its proportion in the public service. Being those clusters: Jewish, Romans, aristocrats, Chinese/Korean among others.
- The surnames with most overrepresentation and underrepresentation in the public service. This means, we take the current proportion of a certain surname, and then it is compared to the proportion of the same surname in the Chilean population (using D1). Finally, we calculate the percent variation, and it is ordered from highest to lowest and from lowest to highest respectively.
- Surnames with highest and lowest average remunerations.
- Regions with the highest and lowest average remunerations.
- Regions with the highest number of public workers.
- Areas of the public service with the highest and lowest average remunerations.
- Areas of the public service with the highest and lowest number of public workers.
- Institutions with the highest and lowest average remunerations.
- Institutions with the highest and lowest number of public workers.

- Rank with the highest and lowest average remunerations.
- Rank with the highest and lowest number of public workers.

4.3.3 Isolation of surnames

The surnames of D2 are inside the attribute of name, in which every name means a full name composed by these elements in the following order: paternal surname, maternal surname, first name, second name. Thus, it is necessary to establish some rules and in some cases smart algorithms implemented in code to isolate each surname (maternal and paternal) from the rest of the name.

In general, the ideal composition of a full name (based on what we figured out in the exploratory analysis) should be a set of words separated by a space string. For example, an ideal full name could be "Gonzalez Jara Luis Miguel", in which "Gonzalez" is the paternal surname, "Jara" the maternal surname, "Luis" the first name and "Miguel" the second name. This assumption implies that for each paternal surname, maternal surname, first name and second name we have only one word (which is not always the case). The ideal full name was tested in the set of full names finding that 92% of the population of D2 meets this format or rule.

The mechanism used was splitting all full names by their spaces, and filtering those which generates (after the splitting function) an array of four words. Finally, we verified that the first two words were Chilean surnames, using database D1. Then, we saved those two words, the first as its paternal surname, and the second one as its paternal surname. In this way, we captured 92% of the both surnames of public workers in database D2.

In the same manner, not all full names meet the rule of the ideal full name, in fact, we still have 8% of the surnames from population D2 that we have to capture and save. For these cases, we build a general approach of an ideal full name, we establish that a surname can be composed by n words separated by $n-1$ spaces. Then, it is necessary to develop an algorithm that can iterate and match successive strings (see Algorithm 1).

Algorithm 1 Surnames matching algorithm. All operations are described below where constant are: S: set of unique Chilean surnames, F: full name workers set.

```
1: procedure SurnameMatching(S, F)
2:     f: is the worker's full name
3:     i: is an auxiliary index for array position
4:     tmp: is a temporary variable for potential surname
5:     elements: is the full name array splitted by spaces Execution
6:     for every f in F do
7:         tmp ← null                                          ▷ Set string
8:         fname ← null                                        ▷ Set string
9:         mname ← null                                        ▷ Set string
10:        i ← 0                                               ▷ Set counter
11:        elements ← Split(f)                                 ▷ Apply split function
12:        while i < elements.length do
13:            tmp ← tmp + space string + elements_i           ▷ Concatenate elements
14:            if tmp in S then                                ▷ Surname found
15:                if fname not null then                      ▷ Father surname found
16:                    fname ← tmp
17:                    tmp ← null                              ▷ Set auxiliary variable
18:                else if mname not null then                 ▷ Mother surname found
19:                    mname ← tmp
20:                    tmp ← null                              ▷ Set auxiliary variable
21:                    break while loop
22:                end if
23:            end if
24:            i ← i + 1                                       ▷ Update counter
25:        end while
26:    end for
27: end procedure
```

In the first place, to reach an efficient matching algorithm for full names, we have to make a splitting operation using the space as target character and separator, then we apply this operation to each full name in the set F. The next step for the algorithm is to initialize a temporary variable which represents the surname that we want to identify. This variable is initialized with the first element of the array (product of the splitting operation).

We enter the loop, then we verify if the temporary variable is a Chilean surname, using the database D1. If the temporary variable belongs to the set of unique surnames S, then it is a paternal surname, because the structure of the full name suggests that the paternal surname is the first element to appear in the sequence. If the temporary variable does not belong to the set of unique surnames S, then the algorithm takes the next element in the array and concatenates two pairs of elements twice, first the temporary variable with a space and then the resulting concatenated string with the next element in the array.

The result will be our new temporary variable.

Later, it verifies again if the new temporary variable belongs to the set of unique surnames S, if it belongs then it is labeled and saved as paternal surname, otherwise the algorithm repeats the concatenate process until the paternal surname is found or when there is no more a next element in the array.

When a paternal surname is found, then the temporary variable is set with the respective next element of the array, and the algorithm repeats the same process used for the paternal surname, but in this case to find the maternal surname. When the maternal surname is found the loop breaks, and the algorithm continues with the next full name in the set F.

In simple words, if we have a person called "San Martín Jara Luis Miguel", then the array product of the splitting operation will contain the following elements: "San", "Martín", "Luis" and "Miguel" (in that order). So when we enter the loop of the algorithm, the temporary variable will be set equal to the string "San", then the algorithm will verify if "San" is a Chilean surname, which is not. Later, the temporary variable "San" will be concatenated with a space and then with the word "Martín", so the new temporary variable will be "San Martín". The algorithm will verify if the word "San Martín" is in the Chilean surnames, which indeed it is. The next step will be saving that word as its paternal surname.

After saving the paternal surname, the temporary variable will be set with the next element of the array which is "Jara", then it will verify if "Jara" is a Chilean surname, which it is. Finally, the algorithm will save "Jara" as its maternal surname, and it will break the loop.

The algorithm was designed and tested for this particular population. In most cases, the paternal surname is found, the same happens with the maternal surname. This reflects a clean database where every worker is characterized by a paternal surname and a maternal surname. Finally, those full names which are not captured by our algorithm (around 0.2% of the population) were fixed manually, by identifying patterns and ways to isolate both surnames.

Isolation of paternal and maternal surnames from the set of worker's full names is a tough task but essential since both surnames are key input to calculate in further sections the metric of paucity or diversity of surnames used in the statistical models.

4.4 Data about active employees in SADP (D3)

As said before, a very interesting cluster that we would like to study is the one in which its members were through a merit-based system (SADP) that claims to be fair and meritocratic by design. Therefore, it is necessary to identify all public workers of database D2 who belong to the SADP.

The data about active employees who were hired through this system is available in the webpage of the National Direction of the Civil Service. This data cannot be downloaded directly, instead we can access the dashboard in the platform of Google Data Studio, then export it to a CSV file.

The obtained information from this source is related to all active public worker hired through the SADP and are characterized by the following attributes:

- *Concurso*: is the code of the public offer published online and on paper by the government.
- *Nivel*: is the level of hierarchy inside the SADP. There are two levels: I and II, "I" being higher than II (more detailed public information was not available).
- *Cargo*: is the name of the charge or role of a public worker.
- *Servicio*: institution that the public worker belongs.
- *Nombre*: first and second name of the public worker.
- *Apellidos*: maternal and paternal surname of the public worker. Both surnames come together in the same attribute, therefore, we have to isolate surnames again.
- *Periodo*: period of the year when the public worker was hired.
- *Inicio*: date of registration of the public worker.
- *Renovación*: indicated if there is a renewal of contract or not.

In this case, we are interested in the data coming from the surnames to characterize the population of the SADP. Paternal and maternal surnames are isolated using the algorithm described in section 5 for the database D2.

	Concurso	Nivel	Cargo	Servicio	Nombres	Apellidos	Periodo	Inicio	Renovación	nombre	Apellido Paterno	Apellido Materno
151	4956	II	Director/a Regional Los Rios	Instituto Nacional de Deportes	FELIPE IGNACIO	MENA VILLAR	Nombrado (primer periodo)	24 feb. 2020	NaN	MENA VILLAR FELIPE IGNACIO	MENA	VILLAR
152	4974	II	Director/a Hospital Carlos Van Buren	Servicio de Salud Valparaiso - San Antonio	JAVIER	DEL RIO VALDOVINOS	Nombrado (primer periodo)	21 feb. 2020	NaN	DEL RIO VALDOVINOS JAVIER	DEL RIO	VALDOVINOS
153	4957	II	Jefe/a Departamento Funcion: Agencia Nacional ...	Instituto Salud Publica	HERIBERTO ENRIQUE	GARCIA ESCORZA	Nombrado (primer periodo)	18 feb. 2020	NaN	GARCIA ESCORZA HERIBERTO ENRIQUE	GARCIA	ESCORZA
154	5056	II	Subdirector/a Administrativo/a Hospital Guille...	Servicio de Salud Concepcion	SERGIO RODRIGO	OSORIO LISPERGUER	Nombrado (primer periodo)	17 feb. 2020	NaN	OSORIO LISPERGUER SERGIO RODRIGO	OSORIO	LISPERGUER
155	5075	II	Director/a Hospital de San Jose del Maipo	Servicio de Salud Metropolitano Suroriente	JAIME ANTONIO	CARVAJAL YANEZ	Nombrado (primer periodo)	17 feb. 2020	NaN	CARVAJAL YANEZ JAIME ANTONIO	CARVAJAL	YANEZ
...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 5: Sample of the database D3. Each row is related to an active employee hired through SADP.

It is important to mention that public employees in this database D3 are also present in database D2, indeed, we matched workers in both datasets in order to analyze this cluster using variables coming from database D2. Furthermore, the records of this database were available until the month of September 2020, therefore, it had to be filtered to get active workers until April 2020, which is the point in time defined.

5 Methodology

The purpose of using the frequency of surnames is to calculate and measure risk of nepotism and elite capture through a homogeneity index (i.e. number of unique surnames). With this KPI, we will be able to determine how likely is to find the current number of unique surnames compared to a simulated number of unique surnames. These simulated numbers are the average of a lot (10^6) of simulations under a random context. Moreover, it is possible to count how many times the simulated number is bigger or smaller than the current number of surnames, therefore, we can calculate a proxy for the statistical significance (p_{value}).

There are multiple models that can be used to determine and figure out events with intervention of random variables, in this case, the model selected for this work will be the Monte Carlo method, because it is simple to implement and very adaptable (see Discussions section). At the same time, we are assuming that randomness might be used to solve deterministic events, thus, we will select random samples repeatedly from a big population (D1). Also, we are assuming a priori by selecting this random sample that the presence of a person in a certain cluster in the public service is not conditioned by the person's surname.

In particular, the objective is to determine in current and random conditions by each cluster in the public service the unique number of surnames, which we called homogeneity index. This means that if we have a set of surnames that contains "Gonzalez", "Rojas", "Larraín" and "Rojas", then the homogeneity index will be equal to 4.

More precisely, we are using this index to determine which units or selected groups may be potential sources of nepotism (a lot of public workers with similar surnames) which means in our study that the current situation (April 2020) has a significant smallest number of unique surnames compared to the simulated context (Monte Carlo's output).

On the contrary, we can identify an overrepresentation (high diversity of surnames) as a proxy for the presence of high socioeconomic groups of the Chilean society which can lead to elite capture. In our research this situation is shown when we observed a significant higher number of unique surnames compared to the simulated context.

Unlike many researches, we do use the whole population distribution, in others words, we use a database of the frequency of surnames aggregated by geographic areas. Detailed records are provided by the IMFD (D1) that help us in the estimation of these simulated values.

Before we run MonteCarlo's simulations, a homogeneity index is built for: (i) each institution, (ii) each rank by region, and (iii) the SADP's cluster. This index represents

the current state of public system in terms of surnames in April 2020. This homogeneity index is called q.

During each iteration of the Monte Carlo simulation, we calculate another homogeneity index this time to represent the simulated situation by our model. This new index is computed by taking random samples from the database D1 (which has the whole universe of Chilean surnames). Then, we count the number of unique surnames of the random sample. For each iteration of the model we gather one random sample, therefore we obtain one simulated homogeneity index called q'_i, where "i" refers to the iteration number.

Consequently, the likelihood of nepotism and elite capture is strongly dependent on the geographic area where it is an institution and the cluster of public workers from a certain rank.

As we said before, we divided our population of public workers in three categories: institutions, ranks and SADP. Then, we have to build 3 three similar but non-identical Monte Carlo models, where each model runs 10^6 times.

5.1 Monte Carlo model by institution

This first simulation intends to establish the average scenario according to the variability or number of surnames of certain public institutions. Therefore, as mentioned above, after running this model we will put special focus on those institutions that:

- Have a considerable low diversity of surnames, i.e. significant lower number of unique surnames, which may indicate high likelihood of nepotism.

- Have a considerably high diversity of surnames, i.e. significant higher number of unique surnames, which may show an unfair representation of the society, in particular, favoring major representation of certain social groups, especially those with high socioeconomic status [37], therefore, this cluster may have a high risk of elite capture.

Before we run the simulation, first we have to build a homogeneity index for each institution. This index called q_o, represents the number of unique surnames in a group of K public workers belonging to the institutions o.

To compute this index, we use the database D2. First, we filter the public workers who belong to the institutions that we want to analyze. Then, we count the number of unique surnames in that set of workers, this gives us the index q_o.

This index will represent the current situation, and it will be compared with the simulated value q'_o, which is the number of unique surnames of the random group of K surnames (using D1) coming from the region where the institutions o is based (using D2).

More precisely, the index q'_o is computed by taking a random group of K surnames from a specific region, and it represents a random sample of K public workers that hypothetically work in the institutions o. To make this first we aggregate the surnames of D1 by region, then we filter that database by the region where the institution is based, so we obtain a temporary table which contains all possible surnames and its frequency for a region k. The region of public institutions is a known value provided by the database D2.

From this temporary table we already have surnames and its frequencies. Later, the algorithm extracts K random surnames from the table. The selection of surnames is not only random, but also it considers the frequency or weight of each surname. In this way, for instance, it will be more likely to select a surname "Gonzalez" (which represents the 0.2 of the population) than a surname "Larraín" (which represents the 0.001 in that same region). The number of unique surnames of this random selection is what we called q'_o.

This value q'_o is simulated 10^6 times.

Algorithm 2 Monte Carlo algorithm for institutions. All operations are described below where constant are: O: set of public institutions, N: 10^6 (iterations), P: database of Chilean surnames.

procedure MonteCarlobyInstitution(G, N, P)
    ▷ o_k: public institution from the region k
    ▷ w_o: set of workers in the institution o
    ▷ n_o: number of workers in the institution o
    ▷ s_o: set of distinct names of workers w_o
    ▷ q_o: size of array s_o
    ▷ w'_o: set of random n workers
    ▷ s'_o: set of distinct names of workers w'_o
    ▷ q'_o: size of array s'_r
    ▷ h_o, l_o, e_o, sum_o: auxiliary variable
    ▷ Execution
    $output \leftarrow dictionary(\{\})$
    for *every* o_k *in* O **do**
        $sum_o \leftarrow 0$ ▷ Set counter
        $h_o \leftarrow 0$ ▷ Set counter
        $l_o \leftarrow 0$ ▷ Set counter
        $e_o \leftarrow 0$ ▷ Set counter
        for $i = 0$ *to* N **do**
            $w_o \leftarrow GetWorkers(o_k)$
            $n_o \leftarrow GetSize(w_o)$
            $s_o \leftarrow GetDistinctSurnames(o_w)$
            $q_o \leftarrow GetSize(s_o)$
            $w'_o \leftarrow RandomChoice(P, n_o, k)$
            $s'_o \leftarrow GetDistinctSurnames(w'_o)$
            $q'_o \leftarrow GetSize(s'_o)$
            if $q_o < q'_o$ **then** ▷ Higher simulated names
                $h_o \leftarrow h_o + 1$
            else if $q_o > q'_o$ **then** ▷ Lower simulated names
                $l_o \leftarrow l_o + 1$
            else ▷ Equal simulated names
                $e_o \leftarrow e_o + 1$
            end if
        end for
        $P_o(q_o < q'_o) \leftarrow h_o/10^6$
        $P_o(q_o < q'_o) \leftarrow l_o/10^6$
        $P_o(q_o = q'_o) \leftarrow e_o/10^6$
        $\bar{q}_o \leftarrow sum_o/10^6$
        $output.add(\{o_k : [\ q_o, \bar{q}_o, P_o(q_o < q'_o), P_o(q_o < q'_o), P_o(q_o = q'_o)\}\ ]\ )$
    end for
end procedure

5.2 Monte Carlo model by regional rank

Before we run the simulation, we must cluster the current public ranks in the database D2, which contains that information. In this case, we build custom ranks for this study based on remunerations, occupations, areas and/or hierarchy inside the organizational chart. Finally, the resulting custom clusters for our study are:

1. *Education* ("Docente", "Docentedirectivo", "Tecnicopedagogico", "No docente de caracter profesional"): personnel who work in education and have a professional degree.

2. *Support* ("No docente de servicios auxiliares", "Auxiliar", "Administrativo", "Administrativos de Salud", "Auxiliares de servicios de Salud" and "Chofer"): personnel who work in government but in the lowest rank. Do not require special qualifications.

3. *Superior Rank*: ("Jefe Superior de Servicio" and "Autoridades de Gobierno"): superior rank which is composed by top managers of public services and cabinet, these positions are generally appointed by the president.

4. *Technical Roles* ("Técnicos de Salud", "Técnico", "Auxiliar Paramédico" and "Técnicos de nivel superior"): personnel who have obtained a technical degree (around 3 or 4-year diploma). Qualifications are necessary for their positions.

5. *Mayors* ("Alcaldes"): Mayors of municipalities.

6. *Professionals* ("Profesional", "Otros Profesionales"): personnel with a professional degree (5 or 6-year diploma), qualifications are required for their positions. This group does not include professional employees that are working in education.

7. *Executive Managers or Directors* ("Directivo"): personnel with a top-tier management role inside public institutions.

8. *Auditors* ("Fiscalizador"): personnel responsible for auditing other public or private institutions.

9. *Health Superior Rank* ("Médicos cirujanos, farmaceuticos, quimicos farmaceuticos, bioquimicos, cirujano dentistas"): health personnel with a professional degree which is considered top-tier such as doctor, dentist, pharmacist, chemist among others.

10. *Managers* ("Jefatura"): personnel with a non-tier management role inside public institutions.

Then, having the data clustered by group, we take each group to split it again, in this case by regions (16 regions), therefore, as a result we will obtain 148 groups. As in the previous simulation, our focus is to determine two possible scenarios:

- Low diversity of surnames compared to the simulated scenario, which may indicate high likelihood of nepotism.
- High diversity of surnames compared to the simulated scenario, which may present major representation of the high socioeconomic groups of the Chilean society [37], then indicates high risk of elite capture.

To run this model, first we have to build a homogeneity index for each of the 148 groups, in which every group index represents the current situation of a certain rank in a certain region (i.e. rank cluster). The index is called $q_{e,r}$, and it represents the current number of unique surnames in the group $G_(e, r)$, which is composed by public workers of the rank e in the region r.

To calculate this index is used in the database D2. First, we divide public workers by the custom cluster described above. Later, these workers are divided again by region, this generates 148 groups. Finally, for each group $G_{e,r}$ we count the number of unique surnames, and that counter is the index $q_{e,r}$.

This index will represent the current situation of each group, and it will be compared with the simulated value $q'_{e,r}$, which is the number of unique surnames in a random sample of K surnames from a region r.

On the other hand, to calculate the index $q'_{e,r}$ we need to select K surnames, where K is the number of public workers of the group $G_{e,r}$). For this step, we will use the database D1, first by aggregating surnames by region, and then by filtering by the region r. These operations generate a table of surname frequencies related to the region r.

After, we take from the generated table K surnames selected randomly. The selection is not only random but also depends on the proportion of each surname in the region. Finally, using the selected group we count the number of unique surnames. This counter is what we called ${q'_{e,r}}'$.

This value $q'_{e,r}$ is simulated 10^6 times.

Algorithm 3 Monte Carlo algorithm for ranks. All operations are described below where constant are: G: set of rank clusters, N: 10^6 (iterations), P: database of Chilean surnames.

procedure MonteCarlobyRanks(G, N, P)
  ▷ $g_{e,r}$: group of rank e in region r
  ▷ $n_{e,r}$: number of workers of rank e in region r
  ▷ $s_{e,r}$: set of distinct names of workers $w_{e,r}$
  ▷ $q_{e,r}$: size of array $s_{e,r}$
  ▷ $w'_{e,r}$: set of random n workers
  ▷ $s'_{e,r}$: set of distinct names of workers $w'_{e,r}$
  ▷ $q'_{e,r}$: size of array $s'_{e,r}$
  ▷ $h_{e,r}$, $l_{e,r}$, $eq_{e,r}$, $sum_{e,r}$: auxiliary variable
  ▷ Execution
  $output \leftarrow dictionary(\{\})$
  for *every* $G_{e,r}$ *in* G **do**
    $sum_{e,r} \leftarrow 0$ ▷ Set counter
    $h_{e,r} \leftarrow 0$ ▷ Set counter
    $l_{e,r} \leftarrow 0$ ▷ Set counter
    $eq_{e,r} \leftarrow 0$ ▷ Set counter
    for $i = 0$ *to* N **do**
      $n_{e,r} \leftarrow GetSize(G_{e,r})$
      $s_{e,r} \leftarrow GetDistinctSurnames(G_{e,r})$
      $q_{e,r} \leftarrow GetSize(s_{e,r})$
      $w'_{e,r} \leftarrow RandomChoice(P, n_{e,r}, r)$
      $s'_{e,r} \leftarrow GetDistinctSurnames(w'_{e,r})$
      $q'_{e,r} \leftarrow GetSize(s'_{e,r})$
      if $q_{e,r} < q'_{e,r}$ **then** ▷ Higher simulated names
        $h_o \leftarrow h_o + 1$
      else if $q_{e,r} > q'_{e,r}$ **then** ▷ Lower simulated names
        $l_{e,r} \leftarrow l_{e,r} + 1$
      else ▷ Equal simulated names
        $eq_{e,r} \leftarrow eq_{e,r} + 1$
      end if
    end for
    $P_{e,r}(q_{e,r} < q'_{e,r}) \leftarrow h_{e,r}/10^6$
    $P_{e,r}(q_{e,r} < q'_{e,r}) \leftarrow l_{e,r}/10^6$
    $P_{e,r}(q_{e,r} = q'_{e,r}) \leftarrow e_{e,r}/10^6$
    $\bar{q}_{e,r} \leftarrow sum_{e,r}/10^6$
    $output.add(\{g_{e,r} : [\ q_{e,r}, \bar{q}_{e,r}, P_{e,r}(q_{e,r} < q'_{e,r}), P_{e,r}(q_{e,r} < q'_{e,r}), P_{e,r}(q_{e,r} = q'_{e,r})\}\ ])$
  end for
end procedure

5.3 Monte Carlo model for SADP

The cluster related to the SADP are public workers who occupy executive or manager roles across the country and were selected by a hiring committee. Therefore, this last model intends to identify in the cluster if they are composed in the same way as high rank clusters or on the contrary, it happens completely the opposite finding more differences than similarities at the moment of comparison with close ranks in terms of public hierarchy. In particular, we want to show if we have high ranks with high risk of nepotism or elite capture, then this situation changes when we are analyzing the national system which claims to be based on merit, equal participation and fair selection, what is called SADP.

This model determines the average scenario in terms of the number of unique surnames in the SADP cluster. Through this computation, we are trying to show if (i) the diversity of surnames is high, indicating that may exist high likelihood of nepotism, (ii) the diversity of surnames is low, indicating that may exist a high risk of elite capture. Both of them contradict the principles of the SADP, and the reason why it was created.

The approach followed for this simulation is very similar to the others, but it has a special difference. Members of each cluster described above are coming from the same region, in particular, in the first case we are analyzing a set of workers from the same organization, and their organization belongs to a certain region while in the second case, we divided ranks by region on purpose. On the other side, SADP members are coming from different regions then we had to adjust our model, in the way that when it selects the random representation of workers, the algorithm must choose workers coming from the 16 regions proportionally.

To run this simulation, first we have to build a homogeneity index for the SADP cluster. The index is called q_S, and it represents the current number of unique surnames in the SADP.

To calculate this index is used D3. Then, we count the number of unique surnames in that database, that counter is the index q_S. This index will represent the current situation of the SADP, and it will be compared with the simulated value q'_S, which is the number of unique surnames in a random sample of K surnames from several regions proportionally.

On the other hand, to calculate the index q'_S we need to select K surnames (where K is the number of public workers of the SADP cluster). For this step, we will use the database D1, then we aggregate surnames by region, this will generate a table of surname frequencies by region. Additionally, we calculate the current number of SADP workers in each region by aggregating the database D3 using the region attribute. In that way, we already know that in the region r are n_r workers.

Later, from the aggregated table coming from D1 we randomly pick n_r surnames for each region r. This selection is not only random but also depends on the proportion of each surname in the region. Finally, from the selection group we count the number of unique surnames. This counter is what we called q'_S.

This value q'_S is simulated 10^6 times.

Algorithm 4 Monte Carlo algorithm for SADP. All operations are described below where constants are: S: set workers of SADP, N: 10^6 (iterations), P: database of Chilean surnames, R: set of Chilean regions.

procedure MonteCarloSADP(S, N, P)
  ▷ s: distinct names in set S
  ▷ q_S: size of array s
  ▷ w_r: SADP workers of the region r
  ▷ n_r: number of SADP workers in the region r
  ▷ W': set of random n workers
  ▷ w'_r: set of random n workers of the region r
  ▷ s': set of distinct names of workers w'
  ▷ q'_o: size of array s'_r
  ▷ h_o, l_o, e_o, sum_o: auxiliary variable
  ▷ Execution
  $output \leftarrow dictionary(\{\})$
  $sum_S \leftarrow 0$ ▷ Set counter
  $h_S \leftarrow 0$ ▷ Set counter
  $l_S \leftarrow 0$ ▷ Set counter
  $e_S \leftarrow 0$ ▷ Set counter
  for $i = 0$ *to* N **do**
    $s \leftarrow GetDistinctSurnames(S)$
    $q_S \leftarrow GetSize(s)$
    $W \leftarrow array()$
    for *every* r *in* R **do**
      $w_r \leftarrow Filterby(S, r)$
      $n_r \leftarrow GetSize(w_r)$
      $w'_r \leftarrow RandomChoice(P, n_r, r)$
      $W.add(w'_r)$
    end for
    $s' \leftarrow GetDistinctSurnames(W)$
    $q'_S \leftarrow GetSize(s')$
    if $q_S < q'_S$ **then** ▷ Higher simulated names
      $h_S \leftarrow h_S + 1$
    else if $q_S > q'_S$ **then** ▷ Lower simulated names
      $l_S \leftarrow l_S + 1$
    else ▷ Equal simulated names
      $e_S \leftarrow e_S + 1$
    end if
  end for
  $P_S(q_S < q'_S) \leftarrow h_S/10^6$
  $P_S(q_S < q'_S) \leftarrow l_S/10^6$
  $P_S(q_S = q'_S) \leftarrow e_S/10^6$
  $\bar{q}_S \leftarrow sum_S/10^6$
  $output.add(\{o_S : [\ q_S, \bar{q}_S, P_S(q_S < q'_S), P_S(q_S < q'_S), P_S(q_S = q'_S)\}\ ]\ )$
end procedure

5.4 Output

For these three models, we calculate four key metrics or KPIs based on the 10^6 iterations made. These four KPIs will be the font of analysis for the next section of Results.

(i) $\bar{q}$: is the average number of unique surnames or average homogeneity index of the 10^6 iterations. It can be computed by adding all simulated indexes and dividing by the total number of iterations. This value is a proxy to determine closeness or distance with the current situation q.

$$\bar{q} = \frac{\sum_{i=1}^{10^6} q'_i}{10^6} \tag{1}$$

(ii) $P(q < q')$: is the probability that the current number of surnames is smaller than the simulated number of surnames. It is computed by counting the number of times when the current number was smaller than the simulated one. Then, this counter is divided by the total number of iterations. This will be our p_{value} used for nepotism risk.

$$P(q < q') = \frac{\sum_{i=1}^{10^6} [q < q'_i]}{10^6} \tag{2}$$

(iii) $P(q > q')$: is the probability that the current number of surnames is bigger than the simulated number of surnames. It is computed by counting the number of times when the current number was bigger than the simulated one. Then, this counter is divided by the total number of iterations.This will be our p_{value} used for elite capture risk.

$$P(q > q') = \frac{\sum_{i=1}^{10^6} [q > q'_i]}{10^6} \tag{3}$$

(iii) $P(q = q')$: is the probability that the current number of surnames is equal to the simulated number of surnames. It is computed by counting the number of times when the current number was equal to the simulated one. Then, this counter is divided by the total number of iterations.

$$P(q = q') = \frac{\sum_{i=1}^{10^6} [q = q'_i]}{10^6} \tag{4}$$

5.5 Logistic regressions

Nowadays, data science is a very hot topic which brings a lot of attention to organizations looking for data driven solutions to their problems. Even though logistic regression

for most data savvy people are known as a machine learning model, it has existed since 1944 in statistics [56].

Developed by Joseph Berkson, logistic regression is a statistical model which uses a logistic function to model, and in general, a binary value representing the dependent variable, which is not always the case since it has more extensions. For this particular case, we will use the logistic regression to calculate, or more precisely, to estimate the parameters of the logistic model.

Through our study, we will use this model four times. The first two regressions are related to institutions. Our goal is to determine features of our data (D1, D2 and D3) that may be related to the risk of nepotism and elite capture in public clusters. Then, we estimate parameters for both logistic regressions (nepotism and elite capture in public institutions).

For the case of nepotism, we label the binary dependent variable as "1" if the institution was identified by the Monte Carlo simulation as very likely to be nepotistic (with 95% confidence interval) and "0" in any other case (see Equation 5). Same criteria we use for risk of elite capture, we label the dependent variable as "1" if the institution was identified by our Monte Carlo simulation as very likely to suffer elite capture, and "0" otherwise (see Equation 6).

$$X_i^n = \begin{cases} 0, & \text{if the institution } i \text{ is very likely to have a high risk of nepotism} \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

$$X_i^e = \begin{cases} 0, & \text{if the institution } i \text{ is very likely to have a high risk of elite capture} \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

Finally, we follow the same approach for rank cluster. Values for the binary dependent variable are set as "1" if the rank cluster is likely to be nepotistic (or to suffer elite capture) according to the p_{values} of the Monte Carlo simulation. We set the value of the dependent variable as "0" in any other case (see Equation 6 and 7 respectively).

$$Y_r^n = \begin{cases} 0, & \text{if the rank cluster } r \text{ is very likely to have a high risk of nepotism} \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

$$Y_r^e = \begin{cases} 0, & \text{if the rank cluster } r \text{ is very likely to have a high risk of elite capture} \\ 1, & \text{otherwise} \end{cases} \tag{8}$$

Independent variables or predictors can be qualitative or quantitative. For this statistical approach, quantitative variables such as number of employees and wages generate no problem. On the contrary, qualitative variables need to be transformed into a quantitative measure. In this way, there are two qualitative variables that we will transform.

The first variable that we want to transform is "region". The assumption here is that it might be a certain tendency going from north to south or from south to north. Since Chile's geography is very diverse due to its long and narrow extension across the Pacific Ocean, we might expect some diversity in terms of population, as well as nepotism or elite capture risks.

In order to establish geographic tendency, we assigned a value of 1 to the most northern region of Chile, the value of 2 to the second most northern region in Chile and so on, until the least northern region (the most southern region) which has a value of 16. In that way, we just have one particular value for each region (see Table 3).

Region	Value
Arica and Parinacota Region	1
Tarapacá Region	2
Antofagasta Region	3
Atacama Region	4
Coquimbo Region	5
Valparaíso Region	6
Metropolitana Region	7
Libertador General Bernardo O'Higgins Region	8
Maule Region	9
Ñuble Region	10
Biobío Region	11
Araucanía Region	12
Los Ríos Region	13
Los Lagos Region	14
Aysén Region	15
Magallanes and Chilean Antarctica Region	16

Table 3: Values assigned to each region in order to prove geographic tendencies when running the logistic regression.

The same happens when we want to assign values to rank, which will be especially important to determine if some specific clusters tend to be more likely to suffer nepotism or elite capture. This last statement could be a potential huge insight for implementing effective public policy. Therefore, we labeled public ranks in a sequential order according to Chilean law [57], with 1 being the lowest and 10 being the highest public rank (see Table 4).

Rank	Value
Support	1
Technical Roles	2
Education	3
Managers	4
Professionals	5
Auditors	6
Health Superior Rank	7
Directors	8
Mayors	9
Superior Rank	10

Table 4: Values assigned to each rank in order to prove ranking tendencies when running the logistic regression.

For modeling these four logistic regression we used the "statsmodels" library, which contains tons of statistical models, and fortunately, it has a logistic regression model compatible with pandas' data frames, which really facilitates the estimation. This library will give us essential metrics such as: correlation, model coefficients, p_{value}, standard error, confidence intervals among others.

6 Results

We have to notice that using the same dataset of workers, we have created three groups: public workers grouped by institutions, by (regional) ranks and workers in the SADP (we excluded temporary workers in all groups). Then, we applied the Monte Carlo algorithm to each group.

The first group is obtained by splitting workers according to the institution which they belong to. The result of this division gives us a total of 577 institutions and each institution is simulated 10^6 times. In the same manner, each institution has only one address which is related to one specific region.

In general, institutions do not seem very representative in terms of surnames, indeed, we have found a total of 274 institutions targeted as very risky, i.e, high likelihood of nepotism or elite capture with a $p_{value} < 0.01$. This overall metric (total of very risky institutions) acquires more importance when we take into account that only 222 institutions were labeled as non-risky at all, this means associated with low risk of nepotism or elite capture, in statistical words, "non-significant" (see Figure 6).

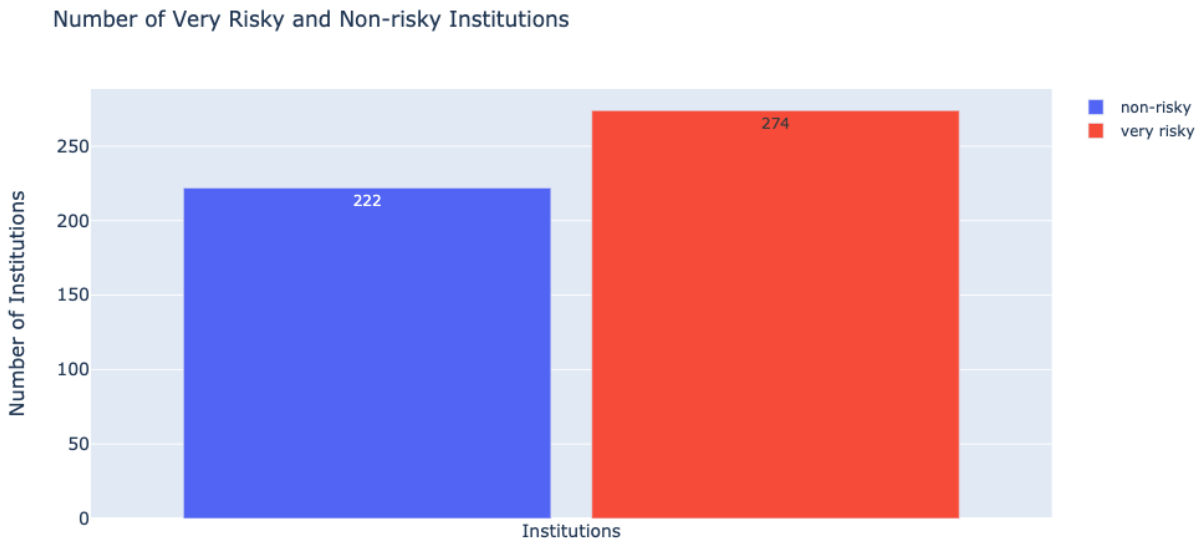

Figure 6: Number of very risky and non-risky institutions. We labeled an institution as very risky if its p_{value} (either of nepotism risk or elite capture risk) is smaller than 0.01. On the other side, we labeled as non-risky if both of its p_{value} (nepotism risk and elite capture risk) are greater than 0.05.

Nonetheless, for this study we will use a threshold of 0.05 for the p_{value} in both cases, nepotism and elite capture risks. Then, the number of risky clusters increases to 355 institutions, which is pretty high considering the total number of institutions in the public

service, in fact, they represent 61.5% of the total institutions in the public service (see Figure 7).

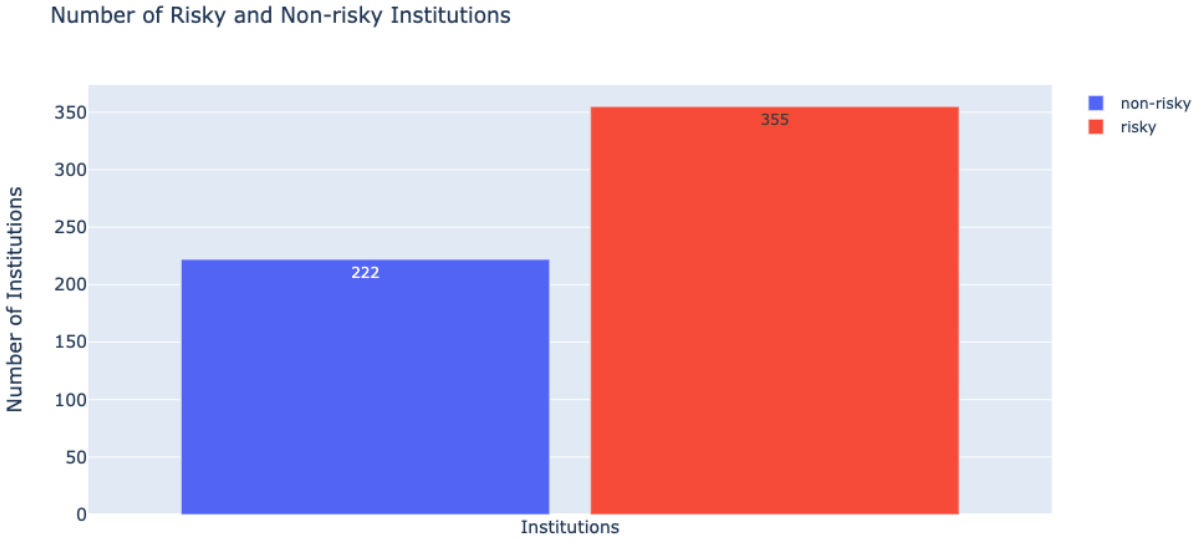

Figure 7: Number of risky and non-risky institutions. We labeled an institution as risky if its p_{value} (either of nepotism risk or elite capture risk) is smaller than 0.05. On the other side, we labeled as non-risky if both of its p_{value} (nepotism risk and elite capture risk) are greater than 0.05.

On the other hand, the second group was obtained by splitting the public workers by their rank, which represents their bureaucratic position in the public hierarchy. In the same way, elite capture (as well as nepotism) is a local phenomena, thus we decided to split these groups again but this time by region.

As a result, we have established a number of 10 clusters by each region (16 regions), which should give us a total of 160 regional rank clusters but since there are some high-ranked ranks that are not present in some regions, we finally obtained 148 rank clusters. Then, each regional rank cluster is simulated 10^6 times using the Monte Carlo algorithm.

The situation for ranks is not very different to the one observed for institutions. In fact, we identified a total of 80 clusters labeled as very risky, in other words, they showed a high risk of nepotism and elite capture with a low p_{value} (less than 0.01). Again, this can be considered as a high value since we identified only 62 clusters which do not present sign related to risk of nepotism or elite capture (see Figure 8).

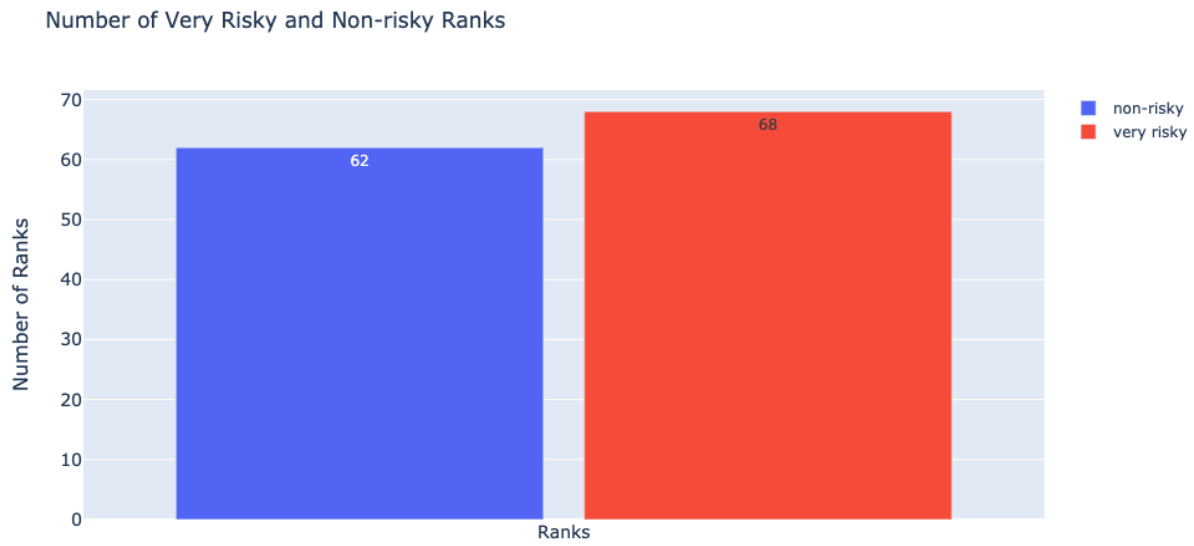

Figure 8: Number of very risky and non-risky rank clusters. We labeled a rank cluster as very risky if its p_{value} (either of nepotism risk or elite capture risk) is smaller than 0.01. On the other side, we labeled as non-risky if both of its p_{value} (nepotism risk and elite capture risk) are greater than 0.05.

As we declared before, our threshold is not 0.01, therefore, in reality local ranks that have been identified as risky are 86. These risky clusters represent 58.1% of the total cluster that we have generated (see Figure 9).

Both institutions and ranks present similar macro risky compositions. Nevertheless, their micro compositions in terms of type of risk are very different. At first sight, rank clusters tend to be more balanced between risk of nepotism and elite capture, meanwhile, the proportion of risk of nepotism tends to be higher in institutions.

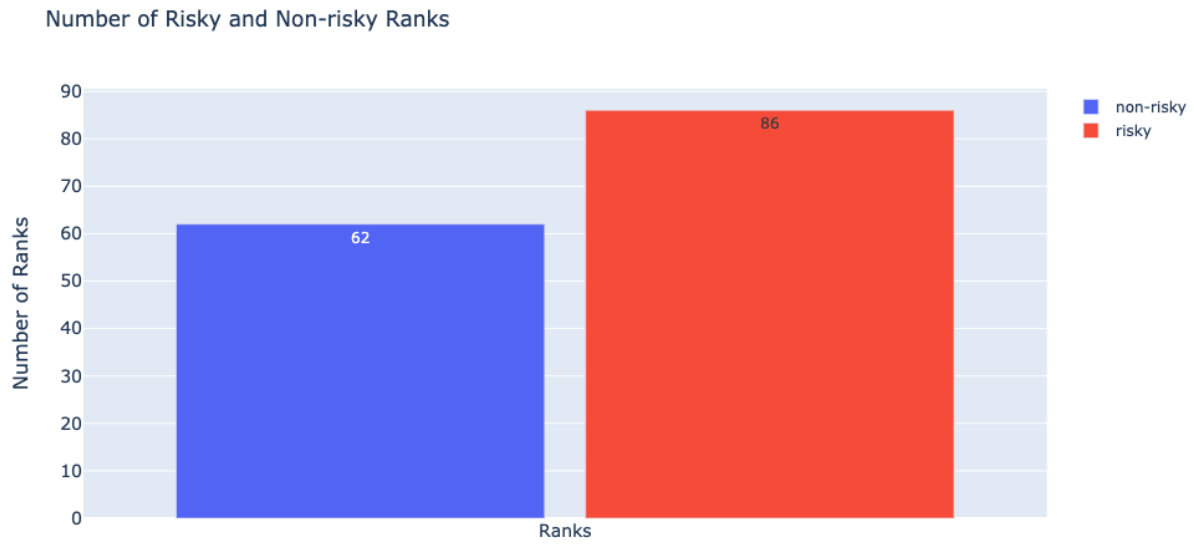

Figure 9: Number of risky and non-risky rank clusters. We labeled a rank cluster as risky if its p_{value} (either of nepotism risk or elite capture risk) is smaller than 0.05. On the other side, we labeled as non-risky if both of its p_{value} (nepotism risk and elite capture risk) are greater than 0.05.

6.1 Institutions

We have described above the risky clusters as generic ones: either associated with risk of nepotism or elite capture. In this section, we intend to go further in the analysis of our results, first by specifying each one. First, we counted the non-risky clusters, identifying 222 out of 577, which accounted for 38.4% of the total of institutions.

On the other hand, we have found 230 clusters (39.8%) that have risk of nepotism, which is a very high number considering the non-risky cluster. Finally, the clusters with risk of elite capture were 125, which is not as high as nepotistic ones but still is about 21.6% of total institutions (see Figure 10).

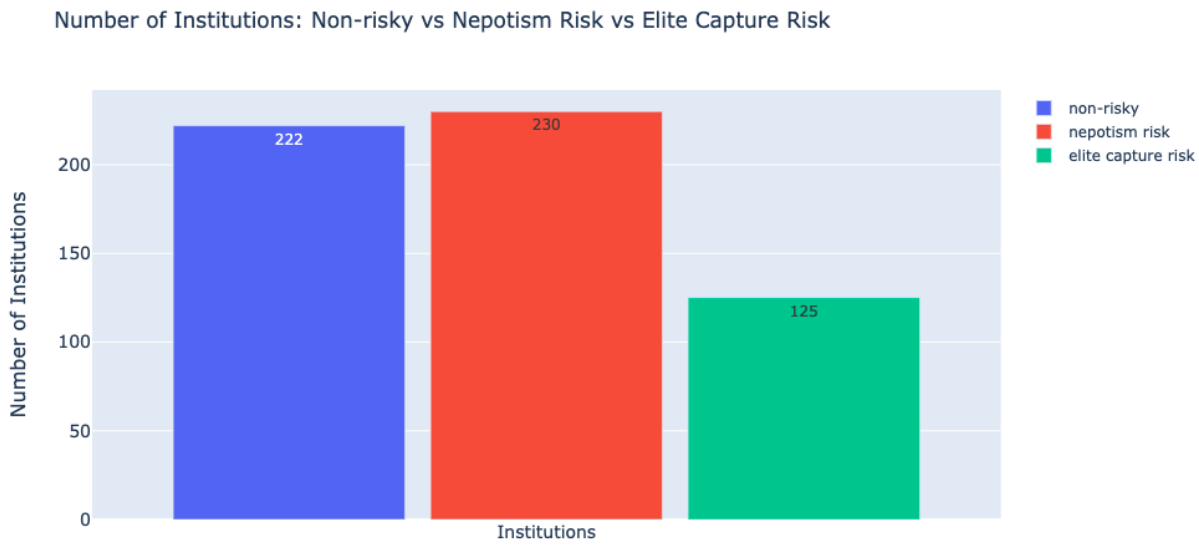

Figure 10: Number of institutions by risk type. We labeled an institution as "nepotism risk" if its p_{value} of nepotism risk is smaller than 0.05. In the same way, an institution is labeled as "elite capture risk" if its p_{value} of elite risk is smaller than 0.05. On the other side, we labeled as non-risky if both of its p_{value} (nepotism risk and elite capture risk) are greater than 0.05.

6.1.1 Region distribution

In the same line, we divide institutions by the region which they belong to. At the moment of counting, by obvious reasons we found that there is a high amount of risky institutions in regions that contain a lot of institutions, such as the Metropolitan Region. Therefore, we decided to express this metric by getting the proportion of institutions with risk of nepotism in a certain region compared to the total number of institutions in that same region (see Figure 11).

Figure 11: Proportion of institutions with nepotism risk by region. Dark colors are associated with a high proportion of institutions with risk of nepotism. Meanwhile, light colors are associated with a low proportion of institutions with risk of nepotism. It is calculated by dividing the number of institutions with risk of nepotism by the total number of institutions.

When analyzing the visualization of the region distribution, we can see that there is not a clear tendency south-north or north-south at first sight. Nonetheless, if we take the top 5 of most risky regions (higher proportion of institutions with risk of nepotism), we will see that 4 of them are southern regions, which may show a nepotistic tendency towards southern regions. Anyway, we will see in further sections if these tendencies are significant for the likelihood of nepotism.

On the other side, we have calculated the proportion for institutions with risk of elite

in each region in the same way as before. Risk of elite capture in institutions is less present than risk of nepotism, but we can notice a certain pattern at first sight, being the southern regions with more proportion of institutions associated with risk of elite capture than the northern ones.

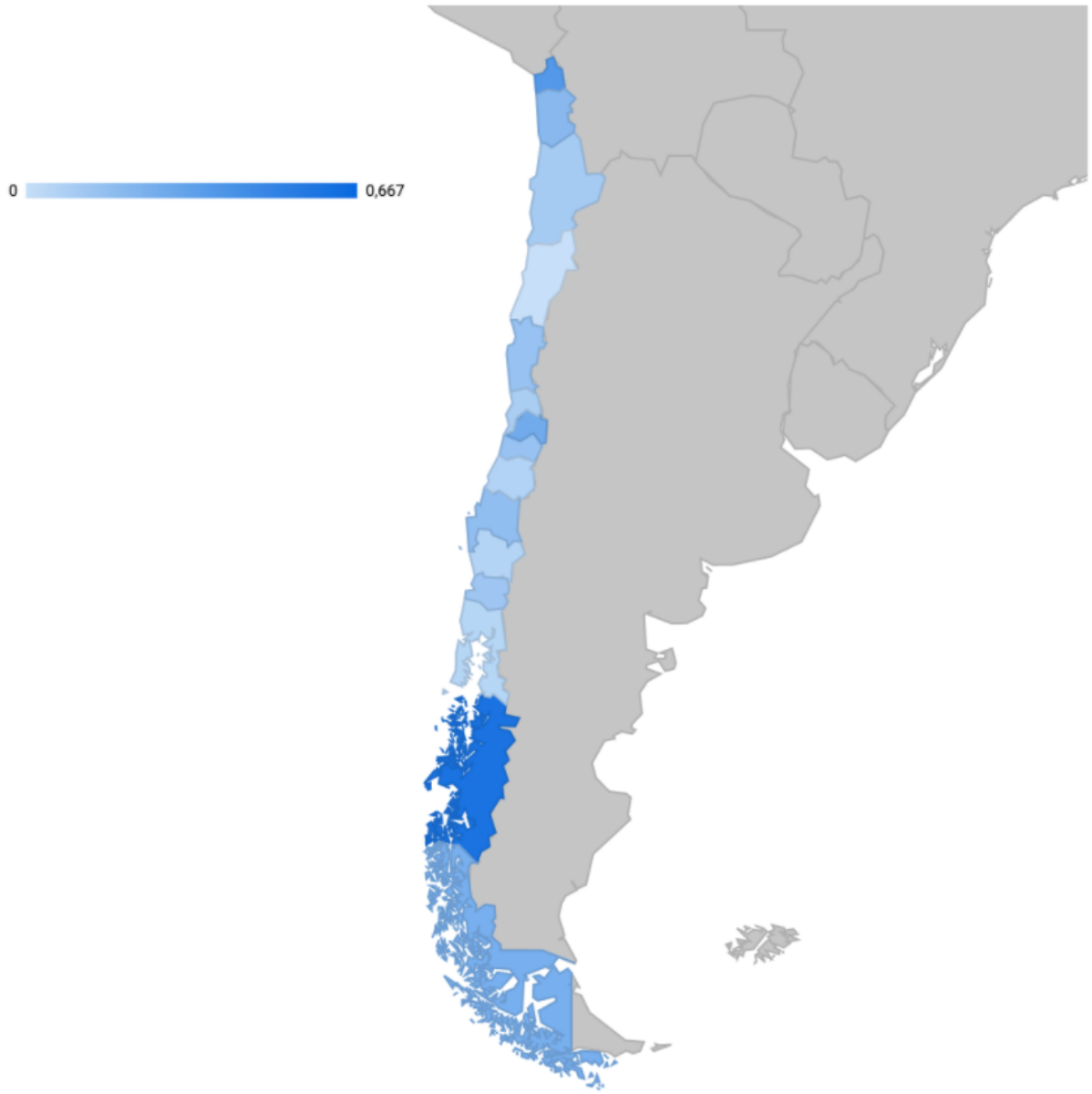

Figure 12: Proportion of institutions with elite capture risk by region. Dark colors are associated with a high proportion of institutions with risk of elite capture. Meanwhile, light colors are associated with a low proportion of institutions with risk of elite capture. It is calculated by dividing the number of institutions with risk of elite capture by the total number of institutions.

Following the same approach, taking the top 5 of most risky regions gives us three southern regions, one northern region and one central region, with the proportion of southern being almost two and a half times higher than their peers. As well as the

nepotistic scenario the tendency of north to south will be tested to identify significance for geographic location of institutions.

6.1.2 Logistic regressions

First, we will analyze the risk of nepotism as a dependent variable. This means that we intend to model the probability of a certain event, in this case, the risk of nepotism in a specific institution. The logistic regression will give us a probability between 0 and 1 of this event.

As we said before, we have defined the dependent variable as risk of nepotism, which in our model is called "label", where each observation gets a value of 1 if the Monte Carlo simulation determines that the institution has risk of nepotism and 0 otherwise.

For the case of nepotism, we used a number of observations of 577 institutions. The result of the logistic regression showed that two independent variables were significant with less than 0.001 of p_value. The two variables were: number of employees and region, both with positive coefficients.

	coef	**std err**	**z**	**P > |z|**	**[0.025**	**0.975]**
Intercept	-1.5158	0.261	-5.805	0.000	-2.028	-1.004
Employees	0.0014	0.000	4.942	0.000	0.001	0.002
Region_ranking	0.0846	0.026	3.201	0.001	0.033	0.136

Table 5: Results of the logistic regression for nepotism risk in institutions.

The fact that both variables had positive coefficients means that the likelihood of nepotism is affected by these two variables in the following way: if we have two institutions in the same conditions, A and B. The number of employees in A is greater than in B, then A has a higher risk of nepotism than B. In the same way, we have found a tendency of increasing likelihood going from north to south. Thus, a southern region will have a higher risk of nepotism than a northern region under the same conditions.

For the second logistic regression, we determine which variables are significantly affecting the likelihood of elite capture in institutions. Again, we defined as our dependent variable the risk of elite capture in a certain institution, being this variable between 0 and 1. As well as for the nepotistic case, we defined the logarithm of the odds as 1 for institutions with risk of elite capture (defined by the Monte Carlo simulation), and as 0 in any other case.

After running the logistic regression, we obtained just one significant variable: wage, which has a negative coefficient. This means that an institution with a lower average wage

of its workers will have a higher risk of elite capture than an institution with a higher average worker wage under the same conditions, and vice versa.

	coef	**std err**	**z**	**P > |z|**	**[0.025**	**0.975]**
Intercept	-3.0554	0.237	-12.888	0.000	-3.520	-2.591
Wage	7.994e-07	9.27e-08	8.621	0.000	6.18e-07	9.81e-07

Table 6: Results of the logistic regression for elite capture risk in institutions.

6.2 Ranks

Rank clusters are special in the sense that they are synthetic groups which do not work together as a whole, but they share a special bond, the level in the public hierarchy, and therefore, a public status.

Even though we are targeting the level of hierarchy as our primary variable to analyze, it would be interesting to see how risky clusters are distributed. By splitting risky clusters, we obtained 55 rank clusters with risk of nepotism (around 37.1%), 43 (29%) with risk of elite capture and 62 as non-risky (see Figure 13).

Figure 13: Number of rank clusters by risk type. We labeled a rank cluster as "nepotism risk" if its p_{value} of nepotism risk is smaller than 0.05. In the same way, a rank cluster is labeled as "elite capture risk" if its p_{value} of elite risk is smaller than 0.05. On the other side, we labeled as non-risky if both of its p_{value} (nepotism risk and elite capture risk) are greater than 0.05.

We should consider that there is more balance among risky types in rank clusters rather than institutions, which were predominantly associated with risk of nepotism. This could

be due to the hierarchy splitting which may produce that some specific members that before were splitting in different institutions, now they are grouped together. This may show new possible dimensions for identifying risk of elite capture.

In the same manner, we aggregate by region the total number of clusters to identify the distribution of the three possibilities or categories: non-risky, risk of nepotism or risk of elite capture (see Figure 14).

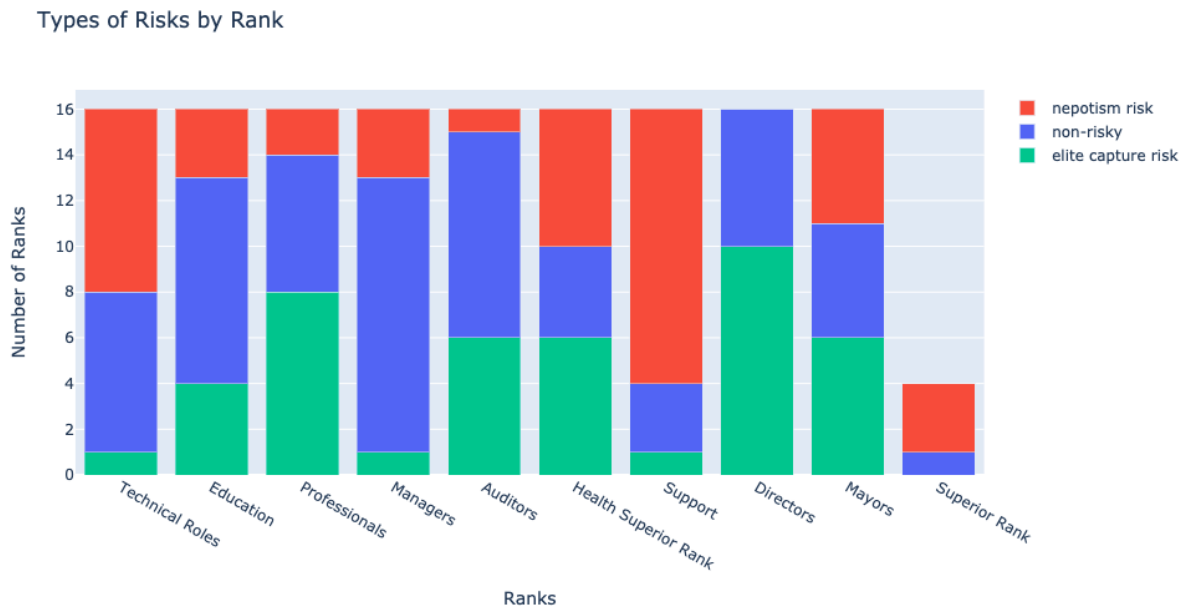

Figure 14: Ranks by their risk types. We followed the same categorization (non-risky, "nepotism risk" and "elite capture risk") by p_{value} described before.

Then, the first thing that we should notice at first sight is that the rank with more rank clusters associated with risk of nepotism is a rank which is very low in the hierarchy, in fact is the lowest (see Figure 15 and 16 respectively).

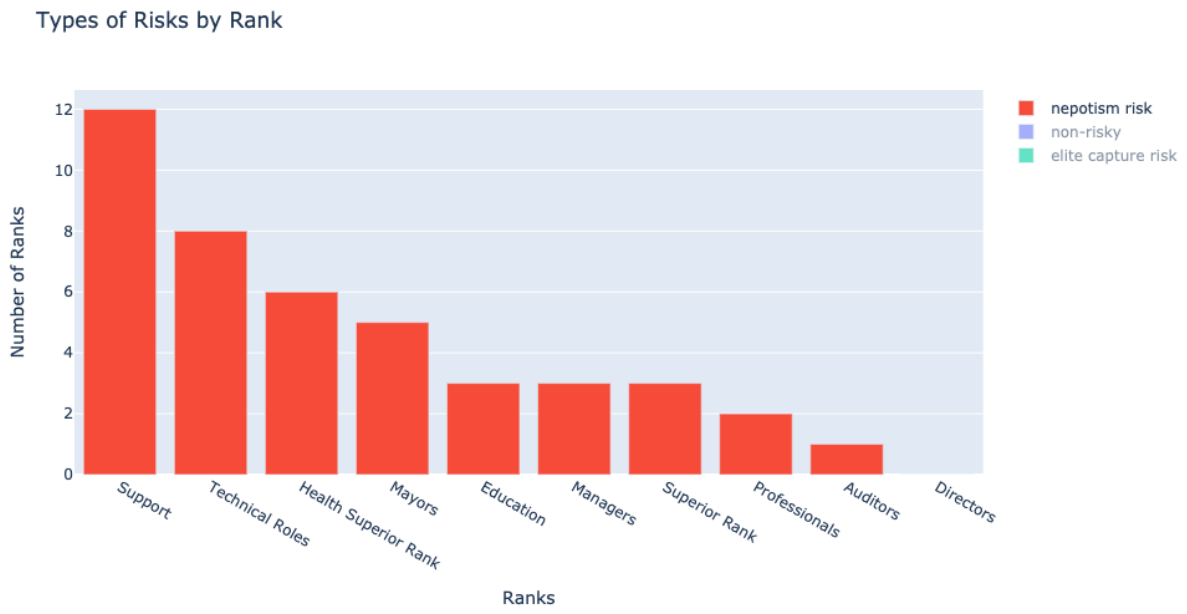

Figure 15: Ranks and their clusters with nepotism risk.

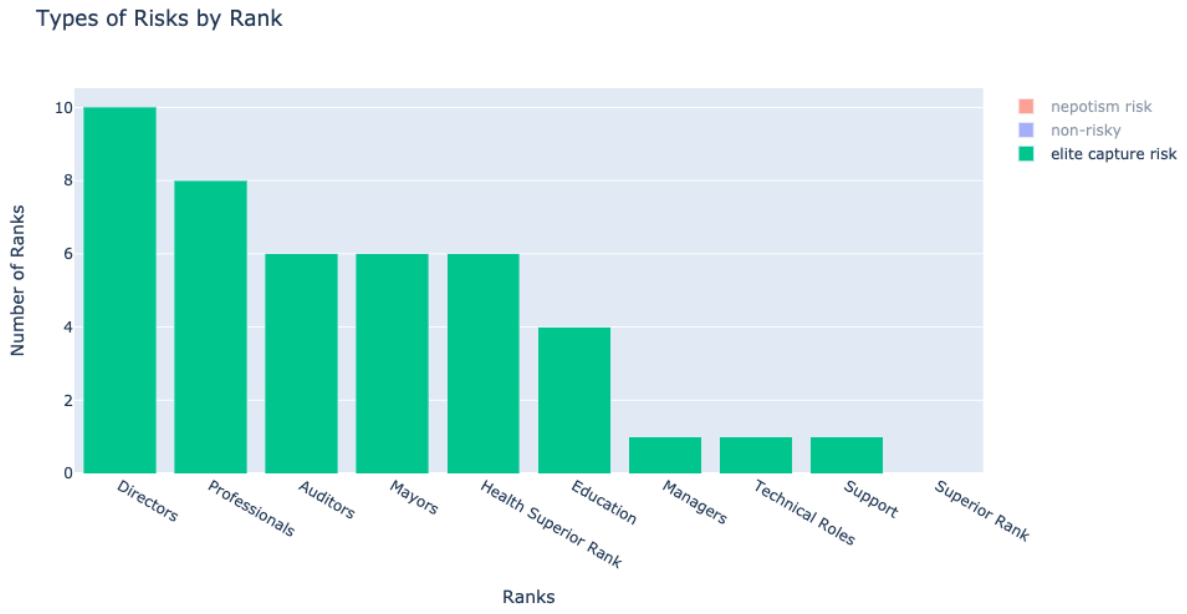

Figure 16: Ranks and their clusters with with elite capture risk.

Meanwhile, the rank that has more clusters associated with risk of elite capture is the "Directives" rank. The fact that we have found the directive rank as the one with more cluster associated with risk of elite capture will be very insightful at the moment of analyzing the SADP group, which falls in the same rank category.

6.2.1 Region distribution

By design, we have divided the ranks by region. Nonetheless, in this case we want to see if there is a pattern in the risk of nepotism by looking at its geographical distribution. In this case, we take the proportion as our main metric to compare since some regions have ten clusters while other regions have only nine.

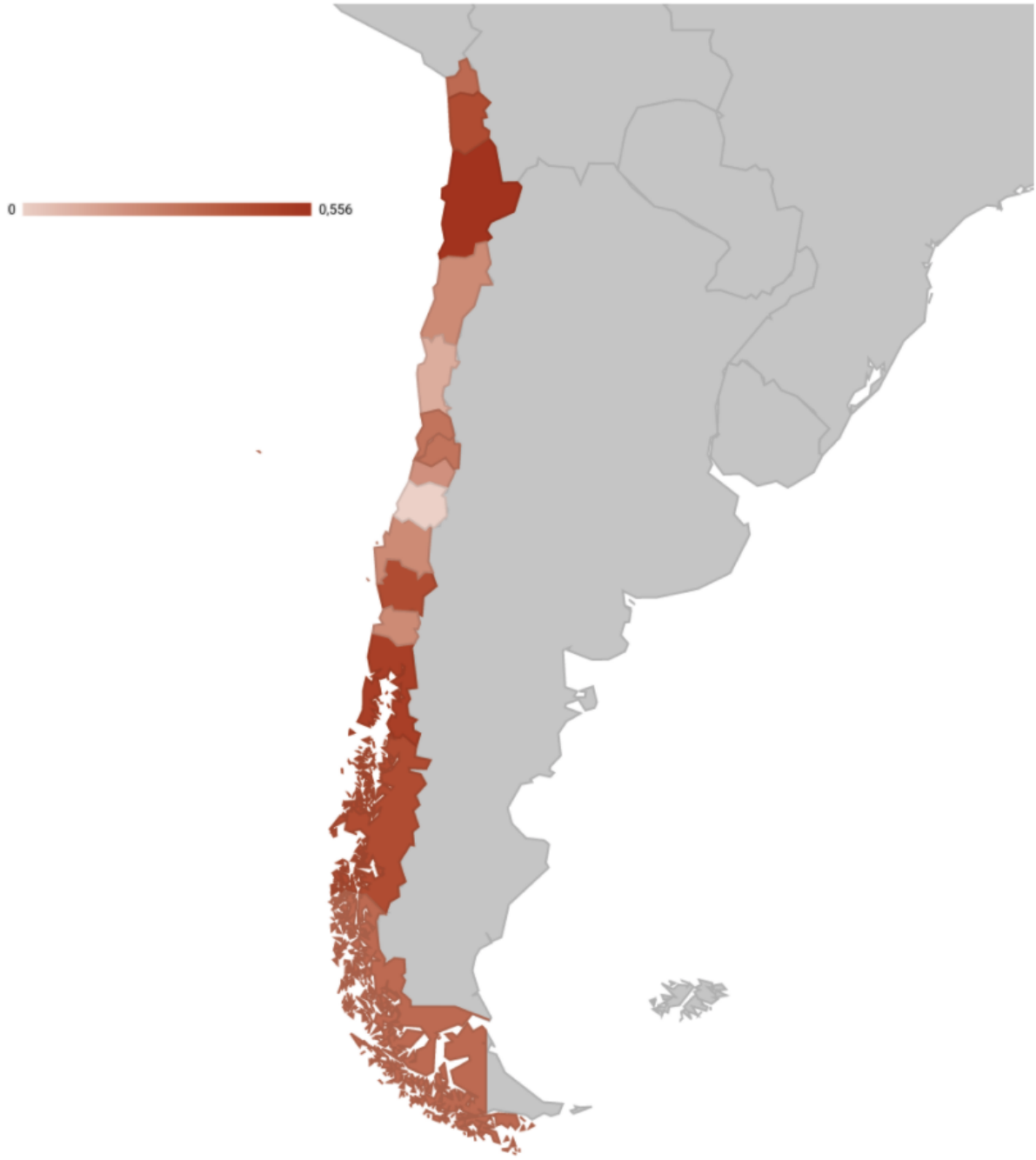

Figure 17: Proportion of rank clusters with nepotism risk by region. Dark colors are associated with a high proportion of rank clusters with risk of nepotism. Meanwhile, light colors are associated with a low proportion of rank clusters with risk of nepotism. It is calculated by dividing the number of rank clusters with risk of nepotism by the total number of rank clusters.

At first glance, we can not notice something relevant in terms of distribution, but we should notice that the extreme north and south may be a hotspot for risk of nepotism (see Figure 17) since the two macro-sectors (group of regions) have around 40-50% of their rank clusters associated with risk of nepotism. Even though there is not a clear pattern looking at regions, this specific topic will be evaluated in the next section when we describe the logistic regression for rank clusters.

For the risk of elite capture, we repeat the steps made for the nepotism risk and region distribution. In this case, we could see a clear geographical pattern, where northern regions have less proportion of rank clusters associated with elite capture risk than southern regions (see Figure 18). That is the first quick insight that we can suppose until now. In the next section it will be statistically proved this geographical variable.

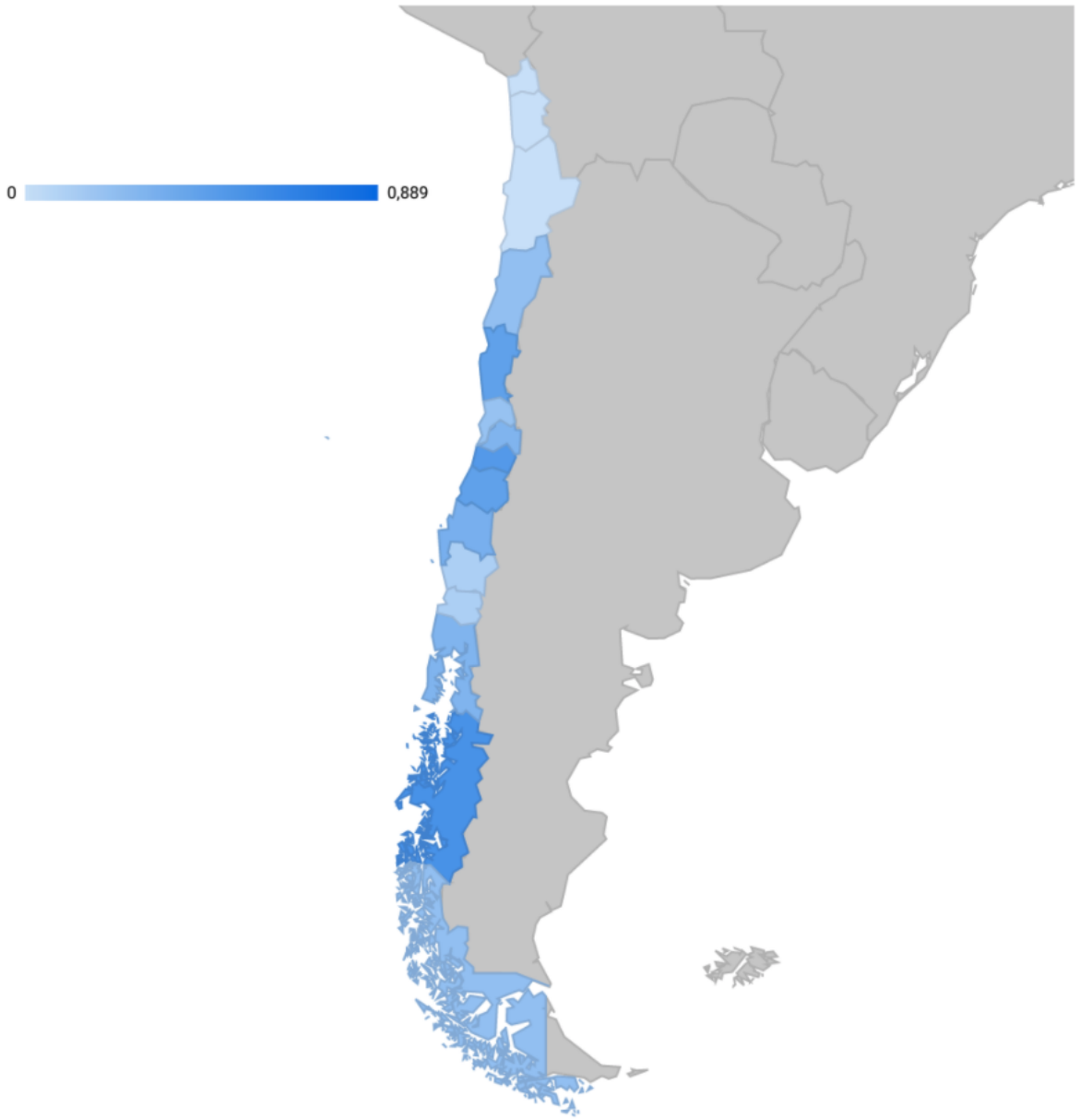

Figure 18: Proportion of institutions with elite capture risk by region. Dark colors are associated with a high proportion of institutions with risk of elite capture. Meanwhile, light colors are associated with a low proportion of institutions with risk of elite capture. It is calculated by dividing the number of institutions with risk of elite capture by the total number of institutions.

Nonetheless, if we take the top 3 of the regions with fewer rank clusters associated with risk of elite capture we found that all three were northern regions, in fact, there are the three most northern regions. On the opposite side, by taking the top 3 of the regions with more rank clusters associated with risk of elite capture we found two southern regions and one central region. This preliminary analysis gives us an idea about the potential of region distribution (north to south) as predictor and its correlation with elite capture risk.

6.2.2 Logistic regressions

In the first regression, we will determine significant variables for nepotism risk. In these terms, using the logistic regression we will be able to establish what variables can influence the likelihood of nepotism in a certain rank.

As we described above in section 5.5, we defined our dependent variable as "label", which means, that it takes the value 1 when it is a rank cluster with risk of nepotism, and 0 otherwise. For nepotism risk, we measured a number of observations of 148 clusters (we had 160 clusters initially, but we removed those with zero members). Then, by iterating through several independent variables such as region, wage and hierarchy ranking among others, we figured out that wage and hierarchy ranking are significant variables with a positive and negative coefficient respectively.

	coef	**std err**	**z**	**P** $>\|z\|$	**[0.025**	**0.975]**
Intercept	0.0436	0.380	0.115	0.909	-0.701	0.788
Cluster_ranking	-0.8958	0.214	-4.192	0.000	-1.315	-0.477
Wage	1.372e-06	3.8e-07	3.614	0.000	6.28e-07	2.12e-06

Table 7: Results of the logistic regression for nepotism risk in rank clusters.

This means that the likelihood of nepotism in a certain rank cluster is higher if the average wage goes up and the hierarchy ranking goes down. On the contrary, the likelihood of nepotism in a certain rank cluster is lower if the average cluster wage goes down and the hierarchy ranking goes up (to a higher rank).

On the other hand, for the second logistic regression which estimates parameters for risk of elite capture, our dependent variable is called "label" again. We modeled these observations by assigning to the logarithm of the odds a value of 1 if the rank cluster has risk of elite capture defined by our Monte Carlo simulation, and 0 otherwise.

In this way, we got three significant variables with a very important p_{value} where each had a value of less than 0.02. The significant independent variables or predictors are: hierarchy ranking, wage and region, possessing a positive, a negative and a positive coefficient respectively. Therefore, the likelihood of elite capture goes up if one or more of these situations happen keeping all underlying conditions: (i) we moved up to a higher rank, (ii) we moved to a cluster in the same rank where the average wage is lower, and (iii) we moved to a southern cluster of the same rank.

	coef	std err	z	P > |z|	[0.025	0.975]
Intercept	-3.5383	0.731	-4.840	0.000	-4.971	-2.106
Cluster_ranking	0.7891	0.204	3.865	0.000	0.389	1.189
Wage	-1.008e-06	3.57e-07	-2.824	0.005	-1.71e-06	-3.08e-07
Region_ranking	0.1076	0.046	2.340	0.019	0.017	0.198

Table 8: Results of the logistic regression for elite capture risk in rank clusters.

6.3 SADP

Finally, we analyzed the set of workers that go through a merit-based selection process. This result will be very critical since it is our only merit-point for comparison, and it may give us some interesting insights.

As we described in the methodology section, we selected a set of random surnames from several regions which depended on the SADP workers distribution across the country. Thus, if hypothetically we have 12, 14 and 15 public employees working on SADP positions in region 1, 2 and 3 respectively. Then, we have to select 12, 14 and 15 random surnames from region 1, 2 and 3 respectively, in order to obtain a representative sample.

Surprisingly, we got that in one million times that we ran the Monte Carlo model, the number of times when the simulated value was lower than the current value was zero. This means that in one million iterations the number of unique surnames taken from the random selection never was lower than the current number of unique surnames in the SADP.

In fact, this gives us a proxy p value of zero, and even though we have found for elite capture risk 99 institutions (17%) and 39 rank clusters (26%) with a p value of zero, this finding is still insightful.

The current number of unique surnames in the SADP is much higher than expected at random. As well as its low percentage of female representation of 32%[51], both situations affect SADP values that at first glance should not include fair or equal representation, which is very critical since we are talking about only top-manager positions, who have a lot of attributions.

In the same way, we should notice that this particular result does not differ from the non-merit director clusters across the country. Indeed, 10 out of 16 director clusters (62.5%) have risk of elite capture (see Figure 19).

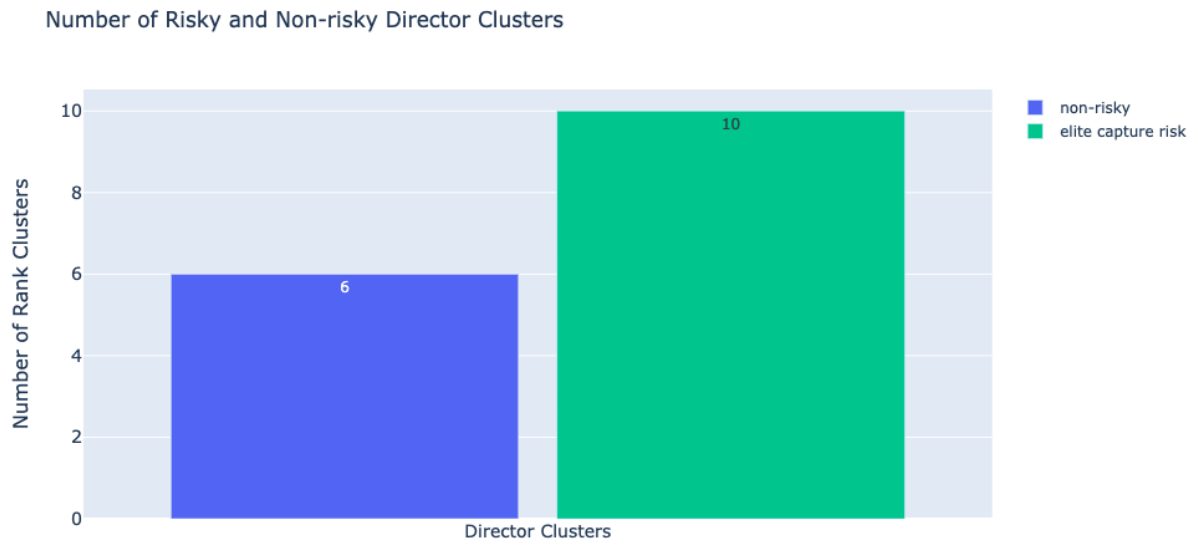

Figure 19: Type of risk of the Director clusters.

This recent discovery should raise a flag at least about the SADP composition and its values. Policymakers should wonder if it would be better to have a representative sample of the population in terms of diversity of surnames.

On the other hand, it is important to verify if this diversity of surnames is intentional or the system selects less frequent surnames due to a self-selection situation where the pool of candidates is already a highly diverse composition.

Finally, we should clarify that since the SADP cluster is the only cluster that follows merit-based procedures in its selection process, and also, is the unique multi-regional cluster in our research, then we did not use it in any of the logistic regressions described before.

7 Discussions

One may think that there are better ways to compute the probability of finding or observing a given number of surnames in a sample. Indeed, this problem has been solved before, but it is not possible to calculate with our current computing local power [58].

Then, a suitable method to find that probability is the Monte Carlo method, in which we hypothetically select random people from certain regions of the population in order to simulate the workers of a given institution or rank. Actually, it makes sense that public workers come from the areas where the institution or rank has its sovereignty. In the same manner, through a random selection we are assuming that applicants and their competencies do not depend on their surnames.

The computation of Monte Carlo simulation is still expensive; indeed, to simulate all 577 institutions it took 332 hours (around 2 weeks). Nonetheless, the algorithm took us little time of coding and development since it is simple to formulate, additionally, the window of time for our research also allows us to take the wait.

The implementation of the Monte Carlo simulation did not require any library, this means that we modeled the simulation end to end. Some technical specialists may argue that is unnecessary since libraries that contain this statistical technique already exist, but in our case did not give us enough flexibility to generate these three similar (but non-identical) Monte Carlo models and that is why we decided to formulate it "manually".

On the other hand, for regressions we found that linear models work well to determine factor parameters that may affect a certain (dependent) variable. But linear models fail in classification problems, which is our kind of problem since we are trying to determine how different variables may affect the likelihood of nepotism or elite capture in a public cluster, and that likelihood is represented mathematically as probabilities.

Linear models cannot output probabilities since they do not work with upper and lower limits, they generate the best hyperplane that fits our values, in other words, output the hyperplane that minimizes the distance between the points (or values) and itself.

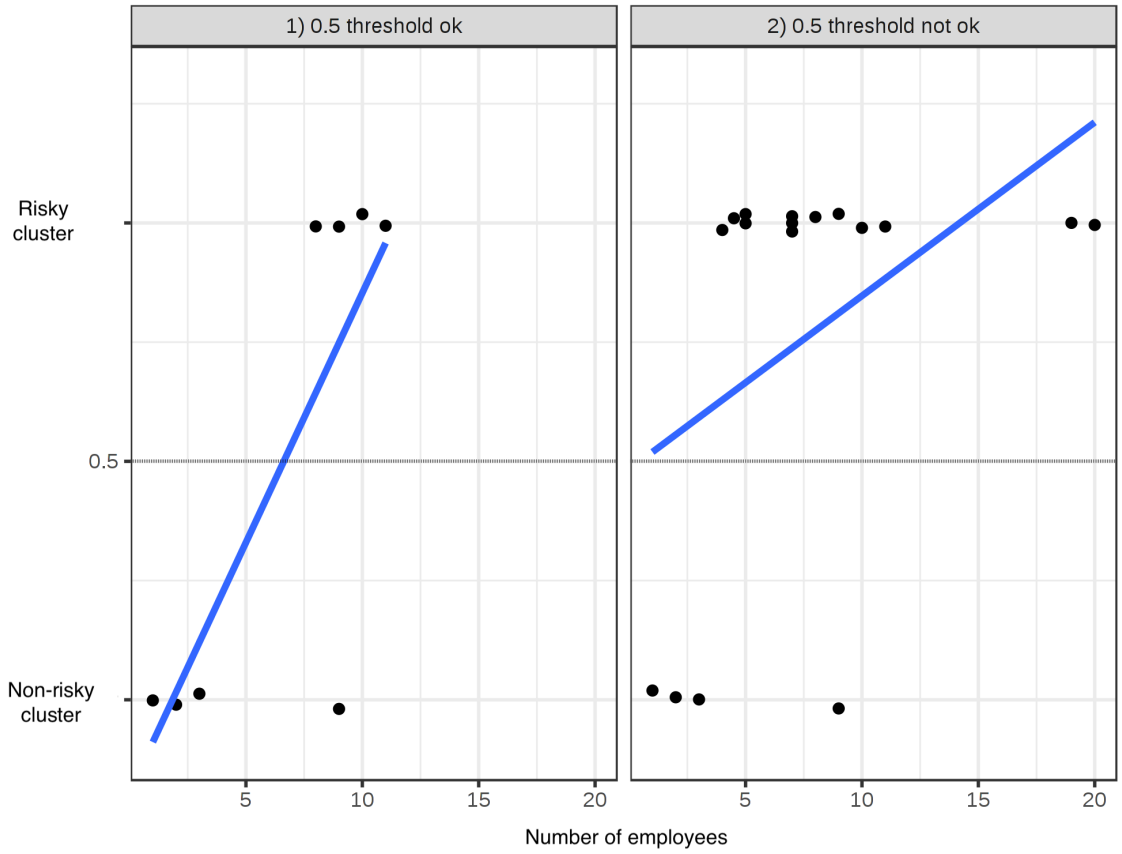

Figure 20: Example of linear regression. In this case, if we have risky clusters labeled as 1 and 0 otherwise, then when we use a linear regression it does not give us the likelihood of the event (being risky or not), in fact, there is no upper and lower limits.

One solution for problem classification can be logistic regressions. Logistic regression helps us to estimate the probability of both risks, of nepotism and elite capture which is based on one or more predictors.

A simple explanation can be shown in a two-dimensional problem, where we have one predictor and one dependent variable. While linear regression allows us to fit a straight line (see Figure 20), logistic regression fits a logistic function that gives us values between 0 and 1 in order to estimate likelihood of categories (see Figure 21).

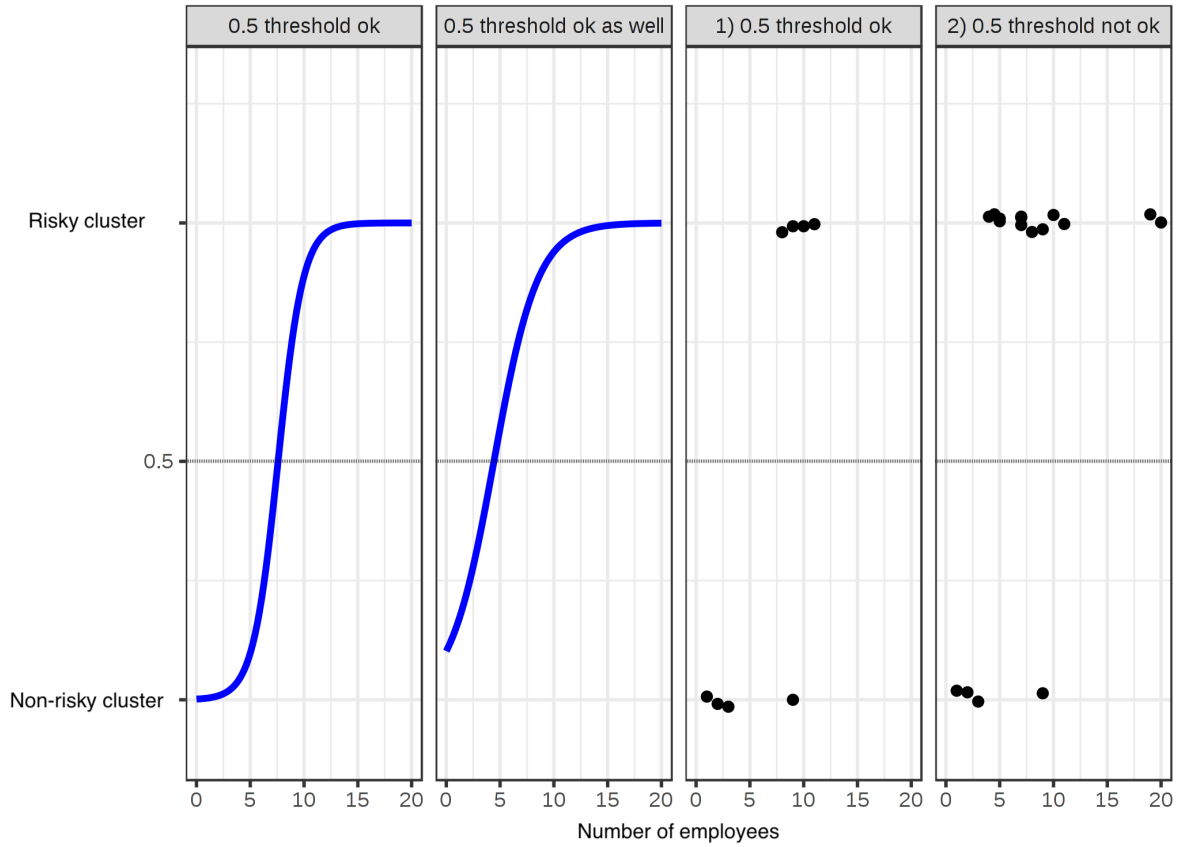

Figure 21: Example of logistic regression. In this case, if we have risky clusters labeled as 1 and 0 otherwise, then when we use a logistic regression it gives us a likelihood (being risky or not), there is upper and lower limits when we increase or decrease the value of the predictor (number of employees).

In previous research, scholars used surnames of the current population to generate their samples. For instance, in this case it would be selecting random surnames from the pool of public workers instead of the whole population, which might bring some bias in the surname analysis.

Therefore, in terms of innovation compared with previous research, our study is one of the first of its type to use the full data of surnames, i.e, data about surname distribution across municipalities in the whole country as the source of sampling, which may produce more accurate results.

By looking at the Chilean context, we could not find any academic research that uses these sources of information (D1, D2, D3) probably because of the size problem of D2, which required high technical knowledge to address it. Nonetheless, we do find some initiatives that use these datasets, specifically D2. One of them is Reguleque.cl which is a search engine that offers all data about public employees since 2013 as a way to improve transparency of public wages, but still it lacks data analysis.

Limitations may arise from our research as well, and one of them is the use of paternal

surnames. For our study we just use the paternal name at the moment of random selection from database D1 (whole universe of Chilean surnames), then our key metric which is the number of unique surnames in a certain sample may change if we take into consideration both surnames. Then, our simulated value and current value might vary a lot, generating different results compared to the proposed in this study.

Also, by taking paternal surnames we are able to identify only the relations of father-child, inter-sibling and most paternal relatives (paternal cousins, aunts and other paternal members) but it leaves outside the scope relationship such as: mother-child, spouses and other close relatives. In the same way, the approach gathers all workers with the same surname even though there may be no linkage between them. Despite all these limitations it is still considered a good proxy to measure the diversity or paucity of surnames.

For the risk of elite capture we also identify some social limitations. The main limitation is the presence of self-selection bias. We know from previous studies that higher income clusters tend to be more diverse in terms of surname. Then, we might expect that, for instance, for higher rank positions the pool of applicants may be already a higher income pool since they eventually at some point got the resources to acquire the necessary competencies.

For example, for SADP positions we might expect elite candidates to apply for it, because these positions required some high level skills that, in the case of Chile, most of the time you have to pay in order to have the opportunity to develop them, which may be the case of bachelor, professional or master degrees.

Summarizing our results, the risk of nepotism prevails over the risk of elite capture in institutions and ranks, but still both are higher than we expected in terms of proportion. On one side, the risk of nepotism is linked to lower ranks, which is something that we may expect because lower ranks require fewer competences than higher ranks, then hiring processes or procedures might be weaker.

On the other side, risk of elite capture is more likely to happen in higher ranks. This is also what we should expect because elite clusters may possess higher qualifications (since they are in an advantageous position to develop them) that are required for higher ranks in the public hierarchy.

Geographic tendencies are something that we do not expect. Actually, at first glance the hypothesis should be that a more isolated region (northern and southern) may be more likely to suffer nepotism and elite capture despite the cluster taken, due to remoteness from the central regions where we often find institutions in charge of the country's control and supervision. Thus, it is a useful insight that helps to justify resources and efforts in southern regions.

Finally, wage correlation was something that we already expected due to previous research at least for risk of nepotism, indeed, some researchers argue that nepotism only exists when public wages are too high [59]. Then, for the case of risk of nepotism makes sense, but for risk of elite capture does not, since lower average wages affects positively and negatively the risk of elite capture in institutions and ranks respectively. Therefore, it is a point in the research that we are willing to discuss further.

8 Conclusions

The approach taken through this study is one of many possible ways. We aim that our work shows how statistical techniques applied to public available data can support policy-makers to address and drive effective public policy.

Therefore, based on the results shown we suggest the following two lines of work for policymakers. Our first suggested line of work is related to hiring processes, where policy-makers have to spend more strong and transparent selection and hiring procedures. From results, we suggest that those efforts and resources should focus on:

- Regions: Aysén, Los Lagos and Ñuble which in average are the regions with higher risk of nepotism and also elite capture.
- The higher and lowest ranks: Inferior rank (risk of nepotism) and Directors (risk of elite capture) including SADP.

In this way we are ensuring that our efforts and resource expenditure will start in the most problematic areas and the most problematic cluster of public workers. Therefore, it would be more easy to measure the impact of these new strong hiring systems.

In parallel, we also need to ensure that job offers that are fulfilled, are also reported to the community in an easy and friendly way. This last statement implies that policy-makers have to rebuild their transparency workflows and user experience through their digital information channels (Transparency Portal).

The first line of work may prevent nepotism in lower ranks and in some institutions. On the contrary, the second line of work is related to assure equal participation and diversity (as we described in the Literature Review section). As an alternative to decrease the risk of elite capture we have to ensure that our pool of applicants for public positions is not already an elite cluster. Then, if we have an elite cluster we should focus on improving the public offering channels to attract a more diverse pool of applicants, in the true sense of diversity. In the same way, since we have identified a lot of high ranks that are very likely to suffer elite capture, then it is also important to implement inclusion and diversity programs, being always aware of bias in the selection process.

For future research it would be interesting to use the data about socio-economic level by each surname. This might help us to verify whether the high clusters are elite clusters or not, and see if the statement "higher the income higher the surname composition" follows previous research.

We have to notice that our study takes only one moment of time: April 2020. And due to the fact that we already have the information of public employees since 2013, another

option for future research is to take that historical data and make a time series analysis with the purpose of tracking the evolution of each institution and rank through time.

And finally, our research sets a precedent about how to solve the problems that bring the analysis of transparency data from the Chilean government. We already explained how we addressed these problems, so we hope that this information opens opportunities for scholars, students and professionals that want to process government data to improve the public service.

9 Bibliography

References

[1] Hayajenh, F., Maghrabi, S. and Al-Dabbagh, T. (1994). "Research note: assessing the effect of nepotism on human resource managers". International Journal of Manpower. 15. 60–7. [DOI:10.1108/EUM0000000003933]

[2] Barozet, E. (2020). Nepotismo, amiguismo y la rabia de los que no son de ningún lote. Retrieved from https://www.ciperchile.cl/2019/08/30/nepotismo-amiguismo-y-la-rabia-de-los-que-no-son-de-ningun-lote/

[3] Jaskiewicz, P., Uhlenbruck, K., Balkin, D. and Reay, T. (2013). Is Nepotism Good or Bad? Types of Nepotism and Implications for Knowledge Management. Family Business Review. 26. 121-139. [DOI:10.1177/0894486512470841]

[4] Darioly, A. & Riggio, R. E. (2014). Nepotism in hiring leaders: Is there stigmatization of relatives?. Swiss Journal of Psychology, 73(4), 243–248.

[5] Burhan, O., van Leeuwen, E. & Scheepers, S. (2020). On the hiring of kin in organizations: Perceived nepotism and its implications for fairness perceptions and the willingness to join an organization. Organizational Behavior and Human Decision Processes. 161, 34-48. Retrieved from https://doi.org/10.1016/j.obhdp.2020.03.012.

[6] Zajac, E. J. & Westphal J. D. (1996). Who shall succeed? How CEO/board preferences and power affect the choice of new CEOs. Academy of Management Journal, 39(1), 64-90. Retrieved from https://doi.org/10.2307/256631

[7] Everett, J., Faber, N. & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. Frontiers in Behavioral Neuroscience, 9, Article 15.

[8] Economía y Negocios de El Mercurio. Retrieved from http://www.economiaynegocios.cl

[9] Kaye, K. (2009). Book Review: Bellow, A. (2003). In Praise of Nepotism: A Natural History. Garden City, NY: Doubleday. Family Business Review, 22(2), 181–183. Retrieved from https://doi.org/10.1177/0894486509333705

[10] Padgett, M. & Morris, K. (2005). Keeping it "All in the Family:" Does Nepotism in the Hiring Process Really Benefit the Beneficiary?. Journal of Leadership & Organizational Studies. 11. [DOI:10.1177/107179190501100205]

[11] Mhatre, K., Riggio, R. & Riggio, H. (2012). Nepotism and leadership. In R. G. Jones (Ed.), SIOP organizational frontier series. Nepotism in organizations. Routledge/Taylor & Francis Group, 171–198.

[12] Gilliland, S. (1993). The Perceived Fairness of Selection Systems: An Organizational Justice Perspective. The Academy of Management Review, 18(4), 694-734. Retrieved from http://www.jstor.org/stable/258595

[13] Lemos, R. & Scur, D. (2012). Management practices, firm ownership, and productivity in Latin America. CAF Working paper, 2012/07, Caracas: CAF. Retrieved from http://scioteca.caf.com/handle/123456789/236

[14] Bennedsen, M., Nielsen, K., Pérez-González, F. & Wolfenzon D. (2007) "Inside the Family Firm: the Role of Families in Succession Decisions and Performance." Quarterly Journal of Economics 122, 2: 647-691.

[15] Rauch, J. & Evans, P. (2000). Bureaucratic Structure and Bureaucratic Performance in Less Developed Countries. Journal of Public Economics. 75. [DOI:10.1016/S0047-2727(99)00044-4]

[16] Slack, C. (2001). Breeding success. MBA Jungle, 82-88.

[17] Arasli, H., Bavik, A. & Ekiz, E. (2006). The Effects of Nepotism on HRM and Psychological Outcomes: The Case of 3, 4, and 5-Star Hotels in Northern Cyprus. 573-587.

[18] Lambert, E., Hogan, N. & Griffin M. (2007). The impact of distributive and procedural justice on correctional staff job stress, job satisfaction, and organizational commitment, Journal of Criminal Justice, Volume 35, Issue 6, 644-656, ISSN 0047-2352. Retrieved from https://doi.org/10.1016/j.jcrimjus.2007.09.001.

[19] Spranger, J. L., Colarelli, S. M., Dimotakis, N., Jacob, A. C. & Arvey, R. D. (2012). Effects of kin density within family-owned businesses. Organizational Behavior and Human Decision Processes, 119(2), 151-162. Retrieved from https://doi.org/10.1016/j.obhdp.2012.07.001

[20] Fafchamps, M. & Labonne, J. (2017). Using Split Samples to Improve Inference on Causal Effects. Political Analysis. 25. 1-18. [DOI 10.1017/pan.2017.22]

[21] Durante, R., Labartino, G. & Perotti, R. (2011). Academic Dynasties: Decentralization and Familism in the Italian Academia.

[22] Bellows, T.J. (2009). Meritocracy and the Singapore Political System. Asian Journal of Political Science, 17, 24 - 44.

[23] Castillo, J., Torres, A., Atria, J. & Maldonado, L. (2019). Meritocracia y desigualdad económica: Percepciones, preferencias e implicancias. Revista Internacional de Sociología, 77(1), e117.

[24] UNDP Global Centre for Public Service Excellence. (2015). Meritocracy for Public Service Excellence.

[25] Young, M. (1958). The Rise of the Meritocracy (2nd ed.). Routledge.

[26] Markovits, D. (2019). The Meritocracy Trap: How America's Foundational Myth Feeds Inequality, Dismantles the Middle Class, and Devours the Elite (illustrated ed.). Penguin Press.

[27] Castilla, E. J. & Benard, S. (2010). The Paradox of Meritocracy in Organizations. Administrative Science Quarterly, 55(4), 543–676.

[28] Castillo, J., Ramírez, S., Atria, J. & Maldonado, L. (2019). Studying Meritocracy in an Unequal Context: Perspectives from Chilean Scholars. Universidad de Talca.

[29] Dahlström, C., Lapuente, V. & Teorell, J. (2012). The Merit of Meritocratization: Politics, Bureaucracy, and the Institutional Deterrents of Corruption. Political Research Quarterly, 65(3), 656–668.

[30] The Northcote-Trevelyan Report. (1954). Public Administration, 32(1), 1–16.

[31] Evans, P. & Rauch, J. (1999). Bureaucracy and Growth: A Cross-National Analysis of the Effects of "Weberian" State Structures on Economic Growth. American Sociological Review, 64(5), 748–765.

[32] McCourt, W. (2007). The Merit System and Integrity in the Public Service. Development Economics and Public Policy Working Paper Series. Paper №20: 1–13. Institute for Development Policy and Management, University of Manchester.

[33] World Bank (1997). World Development Report 1997: The State in a Changing World. New York: Oxford University Press.

[34] Hunt, V., Layton, D. & Prince, S. (2015). Diversity matters. McKinsey.

[35] Castilho, P., Callegaro, H. & Szwarcwald, M. (2020). Diversity matters Latin America: Why diverse companies are healthier, happier and more profitable. McKinsey.

[36] OECD (2009). Fostering Diversity in the Public Service. OECD Paris.

[37] Bro, N. & Mendoza, M. (2021). Surname affinity in Santiago, Chile: A network-based approach that uncovers urban segregation. PLoS ONE 16(1): e0244372. Retrieved from https://doi.org/10.1371/journal.pone.0244372

[38] Gangadharan, L., Jain, T., Maitra, P. & Vecci J. (2014). Impact of Elite Capture on the Provision of Public Services. IGC.

[39] Useem, M. (1979). The Social Organization of the American Business Elite and Participation of Corporation Directors in the Governance of American Institutions. American Sociological Review, 44(4), 553-572. Retrieved from http://www.jstor.org/stable/2094587

[40] Bassett, K. (1996). Partnerships, Business Elites and Urban Politics: New Forms of Governance in an English City? Urban Studies, 33(3), 539–555. Retrieved from https://doi.org/10.1080/00420989650011906

[41] Farazmand, A. (1999). Globalization and Public Administration. Public Administration Review, 59(6), 509-522. [DOI:10.2307/3110299]

[42] Faguet, J., Fabio Sánchez, F. & Villaveces, M. (2020). The perversion of public land distribution by landed elites: Power, inequality and development in Colombia, World Development, Volume 136, 105036, ISSN 0305-750X. Retrieved from https://doi.org/10.1016/j.worlddev.2020.105036.

[43] Platteau, J. , Somville, V. & Wahhaj, Z. (2014). Elite capture through information distortion: A theoretical essay, Journal of Development Economics, Volume 106, 250-263, ISSN 0304-3878. Retrieved from https://doi.org/10.1016/j.jdeveco.2013.10.002.

[44] Bjørnskov, C. (2010). Do elites benefit from democracy and foreign aid in developing countries?. Journal of Development Economics. Volume 92, Issue 2, 115-12. ISSN 0304-3878. Retrieved from https://doi.org/10.1016/j.jdeveco.2009.03.001.

[45] Mookherjee, D. & Bardhan, P. (2005). Decentralization, Corruption And Government Accountability: An Overview. International Handbook on the Economics of Corruption.

[46] International Political Science Abstracts. (2020). International Political Science Abstracts, 70(3), 311–478. Retrieved from https://doi.org/10.1177/0146645320931410

[47] Barron, P. & Clark, S. (2006). Decentralizing inequality? Center-periphery relations, local governance, and conflict in Aceh.

[48] Platteau, J. & Gaspart, F. (2003). The Risk of Resource Misappropriation in Community-Driven Development. World Development. 31. 1687-1703. [DOI:10.1016/S0305-750X(03)00138-4]

[49] Persha L. & Andersson K. (2014). Elite capture risk and mitigation in decentralized forest governance regimes. Global Environmental Change. Volume 24. 265-276. ISSN 0959-3780. Retrieved from https://doi.org/10.1016/j.gloenvcha.2013.12.005.

[50] Servicio Civil (2020). Encuesta Nacional de Funcionarios en Chile: Evidencia para un servicio público más motivado, satisfecho, comprometido y ético.

[51] Servicio Civil. (2020). Página web del Sistema de Alta Dirección Pública. Alta Dirección Pública. Retrieved from https://adp.serviciocivil.cl/concursos-spl/opencms/portada.html

[52] Allesina, S. (2011) Measuring Nepotism through Shared Last Names: The Case of Italian Academia. PLoS ONE 6(8): e21160. [DOI:10.1371/journal.pone.0021160]

[53] Crabtree, J. & Durand, F. (2017). Peru Elite Power and Political Capture. London: Zed Books, 2017. 208 pp. ISBN: 978-1-78360-903-1.

[54] América Transparente (2020). Reguleque. Reguleque.cl. Retrieved from https://reguleque.cl/

[55] Dask (2020). Dask: Scalable analytics in Python. Retrieved from https://dask.org/

[56] Molnar, C. (2021). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Retrieved from https://christophm.github.io/interpretable-ml-book/index.html

[57] Dirección de Presupuestos del Ministerio de Hacienda. (2011). Estadísticas de Recursos Humanos del Sector Público 2001-2010. Retrieved from http://www.dipres.gob.cl/598/articles-82274_doc_pdf.pdf

[58] Walton, G. (1986) The Number of Observed Classes from a Multiple Hypergeometric Distribution, Journal of the American Statistical Association, 81:393, 169-171. [DOI:10.1080/01621459.1986.10478254]

[59] Chassamboulli, A. & Gomes, P. (2019). Jumping the queue: nepotism and public-sector pay.

[60] Musgrave, M. & Wong, S. (2016). Towards a More Nuanced Theory of Elite Capture in Development Projects. The Importance of Context and Theories of Power. Journal of Sustainable Development. 9. 87. [DOI:10.5539/jsd.v9n3p87]

[61] Alatas, V., Banerjee A., Hanna R., Olken, B., Purnamasari, R. & Wai-Poi, M. (2019). Does Elite Capture Matter? Local Elites and Targeted Welfare Programs in Indonesia. AEA Papers and Proceedings, 109: 334-39. [DOI:10.1257/pandp.20191047]

10 Annexes

10.1 Summary

Type of risk	Institutions
non-risky (NS)	173
high risk of nepotism (***)	162
high risk of elite capture (***)	112
risk of nepotism (**)	68
low risk of nepotism (*)	37
risk of elite capture (**)	13
low risk of elite capture(*)	12

Table 9: Summary of institutions. Significance levels "***" ($p_{value} < 0.01$), "**" ($p_{value} < 0.05$), "*" ($p_{value} < 0.1$), NS ($p_{value} > 0.1$).

Type of risk	Clusters
non-risky (NS)	44
high risk of elite capture	40
high risk of nepotism (***)	28
risk of nepotism (**)	15
low risk of elite capture(*)	9
low risk of nepotism (*)	9
risk of elite capture (**)	3

Table 10: Summary of ranks. Significance levels "***" ($p_{value} < 0.01$), "**" ($p_{value} < 0.05$), "*" ($p_{value} < 0.1$), NS ($p_{value} > 0.1$).

Region	Total	N	E
Metropolitana Region	203	61	55
Antofagasta Region	10	6	1
Arica and Parinacota Region	8	0	3
Atacama Region	10	2	0
Aysén Region	15	1	9
Coquimbo Region	20	6	3
Araucanía Region	34	25	2
Los Lagos Region	41	22	2
Los Ríos Region	16	11	2
Magallanes and Chilean Antarctica Region	16	7	4
Tarapacá Region	10	3	2
Valparaíso Region	45	24	4
Ñuble Region	24	1	16
Biobío Region	41	21	7
Libertador General Bernardo O'Higgins Region	40	21	6
Maule Region	30	16	2

Table 11: Summary of institutions by region. Institutions are labeled as N (risk of nepotism) and E (risk of elite capture).

Region	Total	N	E
Metropolitana Region	10	3	3
Antofagasta Region	9	5	0
Arica and Parinacota Region	9	3	0
Atacama Region	9	2	2
Aysén Region	9	4	5
Coquimbo Region	9	1	4
Araucanía Region	9	4	1
Los Lagos Region	10	5	3
Los Ríos Region	9	2	1
Magallanes and Chilean Antarctica Region	9	3	2
Ñuble Region	9	0	8
Libertador General Bernardo O'Higgins Region	10	2	5
Tarapacá Region	9	4	0
Valparaíso Region	10	3	2
Biobío Region	9	2	3
Maule Region	9	0	4

Table 12: Summary of rank clusters by region. Institutions are labeled as N (risk of nepotism) and E (risk of elite capture).