Confounder Parsing for Text Matching

Impact of Confounder Parsing Strategies on Covariate Balance for Text Matching in Service of Causal Inference

Master's thesis in Computer science and engineering

Jannes Reichl
Johan Rönkkö

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2021

Master's thesis 2021

Confounder Parsing for Text Matching

Impact of Confounder Parsing Strategies on Covariate Balance for Text Matching in Service of Causal Inference

Jannes Reichl

Johan Rönkkö

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2021

Confounder Parsing for Text Matching
Impact of Confounder Parsing Strategies on Covariate Balance for Text Matching in Service of Causal Inference
Jannes Reichl
Johan Rönkkö

Supervisor: Adel Daoud, CSE
Examiner: Richard Johansson, CSE

Master's Thesis 2021
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2021

Confounder Parsing for Text Matching
Impact of Confounder Parsing Strategies on Covariate Balance for Text Matching in Service of Causal Inference
Jannes Reichl, Johan Rönkkö
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

In observational studies for policy evaluation, matching is used in service of causal inference to simulate randomization and thus reduce selection bias that might occur when treatment assignment differs systematically. This is done by balancing the distribution of confounding covariates measured before treatments. Matching on numerical covariates has been done for decades. In recent years, matching on textual covariates has gained popularity. By matching on text data, one can potentially observe confounding information that cannot be observed in tabular data. Furthermore, when combined with numerical data, matching on text data can potentially improve the balance of numerical covariates. However, confounder parsing, defined as the process of removing treatment text from documents to only end up with confounding text, is nontrivial in policy evaluation. This is because policy documents come in the form of PDFs and typically vary a lot in terms of quality and layout. There are many different ways in which one could approach confounder parsing and each approach comes with its own trade-offs. We have investigated whether different confounder parsing methods influence covariate balance differently. We applied our methodology to labor issue policies of the International Monetary Fund and measured the impact of these policies on population health. To ensure the relevancy of our inquiry, we also investigated whether text matching improves covariate balance on numerical covariates. We find that the covariate balance of our text matching procedures is relatively unchanged by the different confounder parsing methods. Moreover, text matching within propensity score calipers improves the covariate balance, compared to merely using propensity score matching or matching on text covariates alone. Our results demonstrate that text matching can be valuable in establishing causal inferences in the domain of policy evaluation. In addition, our results also suggest that the flexibility regarding which confounder parsing method researchers can choose among increases.

Keywords: Text matching, confounder parsing, causal inference, covariate balance, International Monetary Fund, policy evaluation.

Acknowledgements

We are grateful for the feedback we have received from our supervisor, Dr Adel Daoud, which has proven to be valuable in critically assessing our progress at many points during the research project and to steer the research design and interpretation of the results in the right direction. Special thanks to Dr. Richard Johansson for assisting us throughout the project to ensure we created a well-structured thesis report. We thank both for the weekly meetings that have proven to be instrumental in keeping us on our track.

Jannes Reichl & Johan Rönkkö, Gothenburg, June 2021

Contents

List of Figures

List of Tables

1 Introduction

Policy evaluation aims to map the effectiveness of policies in achieving their desired outcomes and any additional effects that the policies have. The International Monetary Fund (IMF) is a good example of an impactful organization that has interested policy researchers for decades. The IMF lends money to countries on the condition that such countries implement IMF policies to stabilize their economies (International Monetary Fund, 2020). Research finds that IMF policies generate their desired outcomes, like reducing government budget deficits (Atoyan & Conway, 2006; Dreher & Vaubel, 2004). However, research also shows that IMF policy conditions negatively impact the borrowing countries in vital areas such as health, education, economics and politics (Steinwand & Stone, 2008). These findings are typically a result of comparing the outcomes of interest of countries that were treated (i.e. implemented a policy) with control countries that did not implement a policy. By conducting a study based on a random experiment, one could evaluate policies in an unbiased manner, since the policies would be randomly administered to half the participants (Rubin, 1974). Randomization ensures that researchers can measure the effect of policies regardless of other factors that may make some countries more likely to implement a policy. However, randomly implementing policies would be unethical due to the significant impacts of policies on people's lives and costly due to the size and length of policy implementations.

Therefore, policy evaluators conduct observational studies, where the comparison of treated and control groups is not formed by randomization (Rosenbaum, 1991). Instead, one observes the treatment assignments that naturally occur. As a result, treated and control groups may differ systematically with respect to pre-treatment measures (or covariates) that affect both the likelihood of implementing a policy and the outcome of interest itself. For example, consider evaluating the effect of IMF labor issue policies on child mortality rate. The lower the GDP per capita, the more likely a country is to implement an IMF labor issue policy, due to its higher need for financial aid (Daoud & Reinsberg, 2019). Moreover, a lower GDP per capita also tends to increase the child mortality rate. Comparing treatment and control groups where GDP per capita are different on average, would lead to biased estimates. Such covariates, also known as confounders, are a central concern in observational studies that one needs to control for in order to minimize bias when investigating causal impacts of treatments (Rosenbaum, 1991).

In observational studies, a common technique to control for confounders is by doing matching in service of causal inference (Mozer et al., 2019; Stuart, 2010). Matching is a technique that aims to simulate randomization by equating (or “balancing”) the distribution of confounders in the treated and control groups. In the literature, similarity between these distributions is referred to as "covariate balance" (Stuart, 2010). Covariate balance is improved by matching treatment and control observations that are similar in terms of confounders and potentially pruning those observations for which no adequate match is found. If adequate covariate balance is achieved, treatment assignment can be assumed to be effectively random when conditioned on confounders.

Traditionally, matching has only been done with numerical confounders (Stuart, 2010). In recent years, text data have been incorporated into matching procedures (Egami et al., 2018; Keith et al., 2020; Mozer et al., 2019). With text data, one can potentially reveal confounding information that is unobservable in tabular data. For example, in policy evaluation when using text from policy documents, one might also be able to control for a population’s trust in its government. When making inferences using observational data, common wisdom is to condition on all available covariates that could indicate potential confounding influences (Imbens & Rubin, 2015; Mozer et al., 2019). Furthermore, text matching has to our knowledge not been applied to policy evaluation in the existing literature yet. Natural questions to ask are therefore: 1) whether text matching can improve balancing the distribution of confounders in the treated and control groups, and 2) if text reveals confounding information that is unobservable in tabular data, how this confounding information affects causal effect estimates.

However, policy documents used in text matching typically contain both confounding text and text describing the actual treatment (i.e. the policy that was implemented). Matching treatment and control observations on treatment text might potentially create biased matches. The goal of matching is to match treatment and control units with as similar background as possible before any treatments (Stuart, 2010). Matching on the actual treatment text violates this. Before text matching, these treatment texts would therefore need to be identified and removed to end up with only confounding text to match on. We define this process of removing treatment text from documents as "confounder parsing." Because of the following characteristics of typical policy documents, confounder parsing is nontrivial. First, policy documents generally come in the form of PDF, a format that is known to be hard to parse. Second, a dataset of policy documents can often span over multiple decades. Since the PDF tools have drastically improved over time and the general structure of policy documents in terms of layout and writing typically changes over time, one needs to create a parsing method that covers all these different forms. Moreover, PDFs that date decades back are typically scanned and need optical character recognition (OCR’ing), which further complicates the parsing. The IMF is a great example that fits this description, offering a policy document dataset in the form of PDFs that spans over several decades (in our case 1980-2014). However, many similar text matching use-cases exist for policy evaluation, such as policy documents of the United Nations or the European Union.

There are many different ways in which one could approach confounder parsing and each approach comes with its own trade-offs. For example, one approach might favor removing as much treatment text as possible (including false-positives), while at the same time removing a lot of confounding text. Another approach might favor removing as little confounding text as possible at the cost of not removing treatment text that the parsing approach is uncertain about. A third approach might be very precise in how it removes treatment text, but at the cost of disrupting semantics and word order in the policy documents. Consequently, text matching would be limited to only being able to represent text with a model that does not care about word order, e.g. Bag-of-Words (BoW) models. (Discussed in detail in section 3.3.2.) This study has a strong methodological focus and is less concerned with the causal effects and their implications. In the present study, we aim to investigate how different confounder parsing approaches influence covariate balance when applying text matching in policy evaluation. Hence, our core research question is the following:

> **Research question 1:** For text matching in service of causal inference, do different confounder parsing methods influence covariate balance differently?

Assuming that the choice of confounder parsing method does not matter, researchers can choose the confounder parsing method that suits their research best, without having to be too concerned with the impacts of the confounder parsing design on covariate balance. In order to do this investigation, we use IMF data in an application where the effect of labor issue policies is measured on health outcomes. If text matching does not improve covariate balance, the first research question would be rendered useless, as there would be no point in using text matching. Hence, because we use IMF data, we also address the following question:

> **Research question 2:** Can text matching in service of causal inference improve covariate balance for IMF data, compared to only matching on numerical confounders?

This paper consists of the following sections. First, the theory section will describe the key methodological concepts and elaborate on the functions of the IMF and its impacts. Second, the methodology will lay out the methodological approach and point out its limitations and the potential ethical and risk concerns. Third, the results on the parsing, covariate balance and causal effects will be presented. Fourth, the discussion will offer an interpretation of the results, identify the key limitations of the results and propose recommendations for future research. Finally, the conclusion summarizes the research project.

2 Theory

This chapter first explains the function of the IMF and the impacts of this organization, to give a better understanding of the domain to which we apply our methodology. Thereafter, the methodological definitions are laid out on which our methodology is built.

2.1 International Monetary Fund

The core goal of the IMF is to promote the health of the world economy. To achieve this high-level mission, the IMF focuses its efforts on many sub-goals. Among others, the IMF strives to promote global monetary cooperation, financial stability, international trade, high employment, sustainable economic growth and poverty reduction (International Monetary Fund, 2020). IMF's function can be split into the following three key roles (International Monetary Fund, 2020). First, through economic surveillance the IMF seeks to ensure proper implementation of new policies that are demanded by the IMF in return for IMF loans. Second, the IMF provides technical assistance and training to member countries to improve their economic institutions and policies. Third, the IMF offers loans to countries that need capital to improve their financial balance. IMF loans are often given to stabilize national economies that have been impacted by financial crises, natural disasters or pandemics, such as the covid-19 pandemic. Loans can also be provided to prepare for future crises.

Economically less developed countries are more likely to participate in IMF programs (Steinwand & Stone, 2008). It has been widely established by the academic literature that on average the lower a country's reserves and GDP growth, the higher the number of IMF programs that a country is participating in and the higher the number of conditions that are attached to these programs (Barro & Lee, 2005; Dreher & Vaubel, 2004; Edwards, 2005). Furthermore, a high debt/GDP ratio and low investment flows also increase the likelihood of IMF program participation (Eichengreen et al., 2008; Przeworski & Vreeland, 2000).

According to Ivanova et al. (2003), seventy percent of IMF programs are suspended at a certain point for non-compliance. Countries fail to successfully implement IMF programs for the following reasons, among others. A defective political environment, with a divided government and ethnic divisions increases the likelihood of program interruption. Furthermore, democratic countries are more likely to successfully implement IMF programs, compared to authoritarian regimes (Stone, 2004). A weakly established rule of law also increases the chances of default (Simmons, 2000). Partially due to such weak socio-political infrastructures, the IMF often fails to impose it's conditions even during program years (Evrensel, 2002).

IMF programs have been found to reduce budget deficits (Atoyan & Conway, 2006; Dreher & Vaubel, 2004), stimulate privatization of state-owned assets (Brune et al., 2004), increase private capital flows (Edwards, 2005) and decrease the likelihood of a sudden stop of capital flows (Eichengreen et al., 2008). Privatization can foster healthy market competition, which can promote innovation, improved efficiency and lower consumer prices. Nevertheless, certain basic human needs, such as health care, tend to need government participation to ensure affordable and high-quality access for all citizens. Hence, whether privatization is a net gain or loss for society depends on the use-case and resulting impacts on public spending and public services.

Despite the contributions of IMF programs, they have been found to also negatively impact countries on economic, political, health and education levels. IMF program participation has been found to reduce GDP growth (Butkiewicz & Yanikkaya, 2005; Eichengreen et al., 2008). Furthermore, Bird and Rowlands (2002) and Jensen (2004) discover that IMF lending can lead to a decrease in foreign direct investments. The IMF has also been criticized for promoting moral hazard and dependency on IMF loans (Goldstein, 2002; Hills et al., 1999). Countries may spend money more easily in ways that are not in the nation's best interest, since a new IMF loan can be requested to deal with the resulting budget deficit.

Daoud et al. (2019) shows that almost 60 percent of the policy conditions they analyzed contain policy measures in line with night-watchman state policy preferences. Such a state model prioritizes individual economic freedom and proposes that governments should only interfere to correct market failures (Daoud et al., 2019). The aim of such an approach is to let the free market take care of poverty, through economic growth. Hence, such policies could possibly reduce the influence of the state and stimulate privatization. This finding is further supported by Nooruddin and Simmons (2006), who find that, while in the absence of IMF programs, democracies allocate larger shares of their budget to public services than non-democracies. This difference disappears in the presence of IMF programs. Hence, IMF programs seem to promote the kind of privatization that result in a net loss for society.

The academic literature has widely established the negative impacts of IMF programs on health and education. Nooruddin and Simmons (2006) discovered that participation in IMF programs leads to decreases in education and health spending. More specifically, IMF policy reforms shrink the fiscal space for investment in health, restraining staff expansion of doctors and nurses (Stubbs et al., 2017). Such restrictions limit the advancement towards universal health coverage. Furthermore, (Daoud et al., 2017) finds that IMF programs reduce the protective effect of parental education on child health. For example, for rural citizens, in the absence of IMF programs, living in a household with educated parents reduces the odds of child malnourishment by 38%; in the presence of IMF programs, this protective effect is reduced to 21%. Moreover, (Forster et al., 2019) finds that IMF programs reduce health system access and increase neonatal mortality. These negative effects can largely be attributed to labor market reforms. Finally, (Stuckler et al., 2008) finds that IMF programs lead to tuberculosis increases.

Our methodological contribution in this study, involving confounder parsing and text matching, makes use of an instance based on the research of Daoud and Reinsberg (2019). They analyzed the impact of IMF policies of four different sectors on health. These sectors are fiscal policy, public sector, privatization and price liberalization. They use public-health expenditure, child vaccination and child mortality as their three outcome variables that serve as a proxy for health of the overall population. For the focus of our research, the causal effects of public sector conditions are of interest to us. These public sector conditions include labor policy conditions, which we will use as treatment in this study. One of the most noteworthy findings of Daoud and Reinsberg (2019) is that IMF policy conditions on public-sector employment seems to decrease vaccination rates. Especially conditions that affect doctors and health workers create this negative impact. The average vaccination rate (a percentage) is 10.97% lower when a maximum number of such policy conditions has been implemented, compared to when the minimum number of these conditions has been implemented. Furthermore, Daoud and Reinsberg (2019) finds that public sector conditions increase child mortality and health expenditure, although insignificantly. While the results are robust, it remains unclear whether key unobserved confounding variables affect these results.

Whereas (Daoud & Reinsberg, 2019) adjust for numerical confounding variables and estimate the Average Treatment Effect (ATE), this study adjusts for both numerical and textual confounding using matching in service of causal inference and estimates the Average Treatment Effect on the Treated (ATT). These concepts will be explained subsequently.

2.2 Causal Inference

We use the Rubin Causal Model (RCM) to build an understanding of the causal inference problem in this study. The RCM is a framework for causal inference formalized mathematically, first given name by Holland (1986) for a series of articles by Rubin (1974, 1975, 1976, 1977, 1978, 1979, 1980) that developed the perspective. The RCM builds on three concepts in order to define causal effects: units, treatments and potential outcomes. A unit is a physical object at a particular point in time, in our case a country given a year. A treatment is an action that can be applied on a unit. In our case, we deal with two types of treatments: either a country has implemented a policy of interest, or a country has not implemented a policy of interest. Associated with each unit are two potential outcomes, one for each type of treatment. It is the value of an outcome variable Y at a future point in time from when the treatments started. For example, the outcome Y could be the poverty rate in 1982, with the two potential poverty rates as outcomes had a country implemented a policy or not in 1980. These potential outcomes are measured at the *same* time. Then, the true unit level causal effect would be the comparison between these two potential outcomes. There are different ways of comparing units. In this study we are only concerned with the difference between potential outcomes.

We now formalize this mathematically, adopting the notation from King et al. (2011). We have a data set of N units. Each unit i is assigned treatment T_i, which takes a value of 1 for units who are exposed to the treatment and 0 for units who are not exposed to the treatment. Let $Y_i(t)$ (for $t = 0, 1$) be the value the outcome variable would take if $T_i = t$. The *true* unit level causal effect for observation i is then defined as:

$$TE_i = Y_i(1) - Y_i(0) \tag{2.1}$$

However, since it is impossible to observe both potential outcomes at the same time for a unit i — only $Y_i(0)$ or $Y_i(1)$ can be observed, but not both — the true causal effect cannot be measured (Rubin, 1974). This is known as the "fundamental problem of causal inference" (Holland, 1986). Instead, we *estimate* causal effects (Rubin, 1974). When estimating the causal effect of a treatment on unit i, we compare unit i with a counterfactual unit j who were not exposed to the treatment, i.e. a control unit. (We say that units with $T_i = 1$ belong to the treatment group and units with $T_i = 0$ belong to the control group.) This control unit j cannot be arbitrary. Consider estimating the effect of implementing IMF labor issue policies in Angola in 2010 on child mortality rate. Comparing the child mortality rate of Angola in 2010 with the child mortality rate of Sweden in 2010, who did not implement any IMF labor issue policies, would be misleading. This is because Angola and Sweden are very different countries. One would instead compare Angola in 2010 with another country in 2010 who are identical, or as similar as possible to Angola in 2010, with respect to the set of most important background covariates (or control variables).

Let X_i denote the vector of covariates *before* unit i has been treated. In the previous example, X_i could consist of variables such as prior IMF programs, reserves and GDP growth. For a treated unit i with only $Y_i(1)$ observed, we would estimate its causal effect by finding a counterfactual control unit j that is identical to i with respect to the background covariates (i.e. such that $X_i = X_j$), or as similar as possible to i (i.e. $X_i \approx X_j$) before treatments and compare their outcomes:

$$TE_i = Y_i(1) - Y_j(0) \tag{2.2}$$

Finding a control unit j that is identical to i with respect to the background covariates is usually not feasible, especially in observational studies. Most studies therefore settle with comparing treated and control units that are similar. In order to achieve good estimations of unobserved potential outcomes, we want to compare treated and control units that are as similar as possible (Stuart, 2010).

So far we have explained causal effects on a unit level. However, most studies are interested in estimating average causal effects of a treatment. The Average Treatment Effect (ATE) is the expected causal effect of a treatment across all units in the population set. It is calculated as:

$$ATE = E[Y(1) - Y(0)] \tag{2.3}$$

In our previous example, the ATE calculates the expected causal effect of implementing an IMF labor issue policy on a country's child mortality rate. In some cases, we are interested in, or are only able to calculate, the expected causal effect of the treatment for units who have been treated. In this case, we calculate the Average Treatment Effect on the Treated (ATT):

$$\begin{aligned} ATT &= E[Y(1) - Y(0)|T = 1] \\ &= E[Y(1)|T = 1] - E[Y(0)|T = 1] \end{aligned} \tag{2.4}$$

In our previous example, as only developing countries participate in IMF programs, the ATT calculates the expected causal effect of implementing an IMF labor issue policy on a developing country's child mortality rate. In this study, as IMF policy documents are only available for countries that have participated in an IMF program, we estimate the ATT and not the ATE.

In order to make unbiased claims from these estimations, the following assumptions must hold.

- Stable Unit Treatment Value Assumption (SUTVA) (Cox, 1958; Rubin, 1980)
- Ignorability assumption (Rosenbaum & Rubin, 1983)
- Overlap assumption (Rosenbaum & Rubin, 1983)

SUTVA relates to how one defines what a treatment is and needs to be accounted for in both randomized experiments and observational studies (Cox, 1958; Rubin, 1980), whereas the ignorability and overlap assumptions relate to treatment assignment and are assumed to hold in randomized experiments as treatment assignment is random and not systematic (Rosenbaum & Rubin, 1983). In observational studies where treatment assignment might differ systematically, we need to control for covariates that impact both the probability of receiving a treatment as well as the outcome (known as confounders) in order to account for the ignorability and the overlap assumptions. We explain these assumptions subsequently.

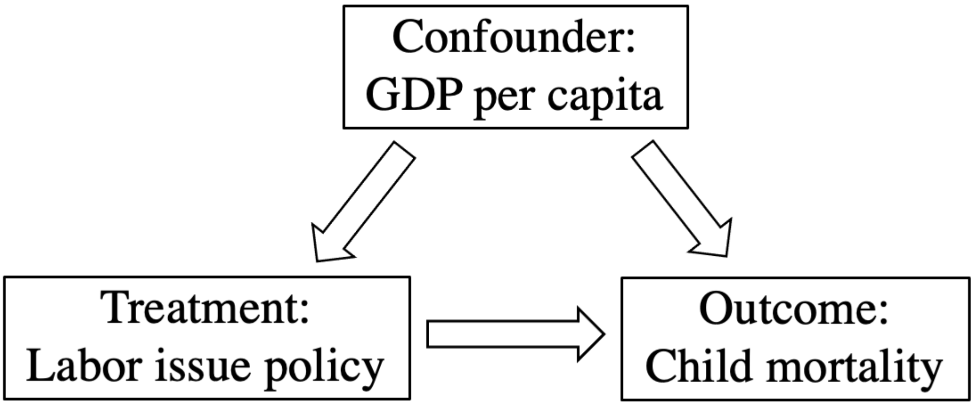
2.2.1 Stable Unit Treatment Value Assumption (SUTVA)

An assumption necessary to estimate causal effects is the Stable Unit Treatment Value Assumption (Cox, 1958; Rubin, 1980). It states that 1) the treatment of one unit does not affect the potential outcome of other units, and 2) that a treatment has the same effect on a unit regardless of how the unit came to be exposed to the treatment. This is necessary in order to ensure that the causal effect for each unit is stable. Violating either of the two aspects of SUTVA might lead to causal inferences that cannot be trusted since estimates of the causal effect are unstable. By "unstable" we mean that there might be multiple versions of a treatment, although it was defined as a single construct, and each version might influence a particular unit in a different way (Schwartz et al., 2012). In the case of IMF policy evaluation, SUTVA might be violated if we are not careful with our definition of what a policy is. We know that all IMF policies are unique in the sense that they are written specifically for a loan, but they can also be categorized into policy area, arrangement type, etc. Depending on the level of abstraction when defining what policy is, estimated causal effects might be more or less stable.

2.2.2 Ignorability Assumption

Unlike randomized experiments, in many observational studies, treatment assignment is not random. Instead, treatment assignment might differ systematically with respect to pre-treatment measures or covariates. Factors other than the treatment may impact (or confound) the outcome variables of interest and consequently bias the estimated causal effects (Rosenbaum, 1991; Rosenbaum & Rubin, 1983). Consider Figure 2.1 below together with the previous example about estimating the expected effect of labor issue policies on a country's child mortality rate. Let's assume that GDP per capita is the only covariate in this example that causes a country to implement a labor issue policy while at the same time impacting the child mortality rate, i.e. a confounder. If the average GDP per capita in the treatment group is different from the average GDP per capita in the control group, it may bias the causal effect estimation. On the other hand, if the average GDP per capita in both groups are the same, we can *ignore* GDP per capita. Hence, there is no dependence between child mortality rate and the labor issue policies and we can estimate the causal effect in an unbiased way, assuming SUTVA and the overlap assumption hold.

Figure 2.1: Confounder relationship with treatment and outcome.



This motivates the ignorability assumption (Rosenbaum & Rubin, 1983), which states that treatment assignment (T) is independent of the potential outcomes

$(Y(0), Y(1))$ given the covariates (X):

$$(Y(0), Y(1)) \perp T|X \tag{2.5}$$

It means that treatment assignment can be assumed to be effectively random when conditioned on covariates. The ignorability assumption is also termed "unconfounded" or "no hidden bias" (Stuart, 2010).

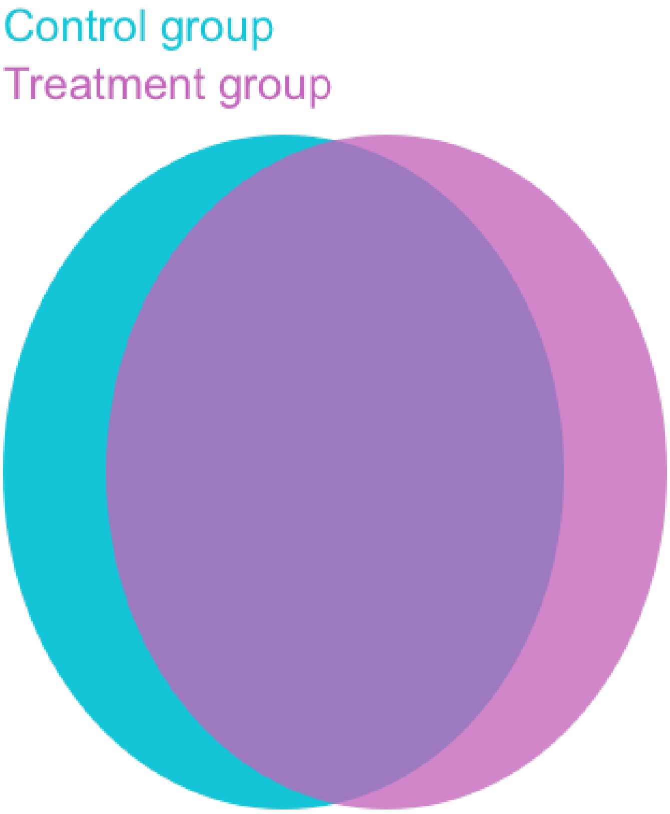
2.2.3 Overlap Assumption

The overlap assumption (Rosenbaum & Rubin, 1983), also known as common support, states that there is a positive probability of being assigned to the treatment group all values of X:

$$0 < P(T = 1|X) < 1 \tag{2.6}$$

Consider Figure 2.2, which depicts an example of lack of complete overlap in covariate distributions across treatment and control groups. The turquoise oval on the left represents the control group, the purple oval on the right represents the treatment group and the dark purple oval in the middle represents the overlap.

Figure 2.2: Example of partial overlap between treatment and control covariate distributions.



In order to make unbiased causal effect estimates, Gelman and Hill (2006) states that one must either "restrict inferences to the region of overlap, or rely on a model to extrapolate outside this region." If one does not handle the lack of overlap, one would end up with a sample consisting of treatment and control units with no counterfactuals, which would yield biased causal effect estimates. Model dependence is a slippery slope as it only shows how well a researcher was able to find a model consistent with prior expectations (Gelman & Hill, 2006). Instead, we want a mechanism to e.g. partition our sample in a way such that we only end up with treatment and control units in the overlapping area.

When both the ignorability and overlap assumptions hold, we say that treatment assignment is strongly ignorable (Rosenbaum & Rubin, 1983). In this case, treatment assignment is unrelated to the covariates, i.e. such that

$$\tilde{p}(X|T=1) = \tilde{p}(X|T=0) \tag{2.7}$$

where $\tilde{p}$ denotes the empirical distribution. To achieve strongly ignorable treatment assignment is the goal of matching in service of causal inference (Stuart, 2010), which is explained subsequently.

2.3 Matching

The goal of matching in service of causal inference is to simulate randomization by equating (or “balancing”) the distribution of confounders in the treated and control groups in a way so that the ignorability and overlap assumptions can be justified (Mozer et al., 2019; Stuart, 2010). It does this by creating matched pairs between treatment and control units that are similar with regards to specified background covariates and prune those treatment and control units for which no adequate match is found. To what degree one can make the ignorability and overlap assumptions hold, depends on the data, what covariates are included in the matching procedure and how one define what an adequate match is. Traditionally, matching has only been done on numerical data. In recent years, text data have been incorporated into matching procedures (Egami et al., 2018; Keith et al., 2020; Mozer et al., 2019). We start by explaining traditional matching and thereafter explain how one can incorporate text into a matching procedure. Stuart (2010) breaks down traditional matching on covariates into four key steps.

1. Define a distance metric to measure how similar a treatment and control unit are with each other, i.e. whether they are a good match or not.
2. Implement a matching procedure that matches treatment and control units with each other, given the distance metric.
3. Assess the quality of the resulting matched sample. Repeat steps 1-2 until a well-matched sample is achieved.
4. Analyze the outcome(s) of interest and estimate treatment effect, given the matched sample in step 3.

These four steps are explained in detail below.

2.3.1 Distance Metric

The first of matching is to select a distance metric. There are two main aspects one need to consider when defining the distance metric to use in matching (Stuart, 2010):

1. what covariates to include, and
2. how to combine those covariates into one distance measure.

Matching aims to simulate randomization and thereby relies on the ignorability assumption, i.e. that there are no unobserved differences between the treatment and control groups (Stuart, 2010). In order to satisfy this assumption, it is therefore important to include all covariates that are known to impact both the probability of being treated and the outcome, i.e. all known confounders (Glazerman et al., 2003; Heckman et al., 1998; Hill et al., 2004; Rubin & Thomas, 1996). For some distance metrics, e.g. propensity score distance explained below, one can be liberal with what covariates to include, as including variables unassociated with treatment assignment will have little influence in the matching procedure (Stuart, 2010). However, excluding an important confounder or including a variable that may have been affected by the treatment, can be very costly in terms of increased bias (Frangakis & Rubin, 2002; Greenland, 2003; Rosenbaum & Rubin, 1984; Stuart, 2010).

There are three types of distance metrics: exact, coarsened exact, and continuous distance (Stuart, 2010). Exact distance considers whether or not two units are identical with respect to their covariates, i.e. $X_i = X_j$. If so, the two units are a match. Coarsened exact puts covariates into bins and measures similarity based on the number of shared bins. In contrast, a continuous distance metric produces a scalar value from the covariates of two units, which represent the similarity between those two units. Popular continuous distance metrics are: Euclidean distance, Mahalanobis distance, cosine distance and (linear) propensity score distance. For continuous distance metrics it is not enough to provide a distance formula, but one is also required to specify a *caliper*. A caliper specifies the maximum allowable distance at which two units must be said to still match. By specifying a caliper, one also discards any treated units for whom there are no control units within a suitable distance, which help satisfy the overlap assumption.

In this study, we use linear propensity score distance when matching on numerical covariates. A propensity score, introduced in (Rosenbaum & Rubin, 1983), summarizes all of the covariates of a unit into one scalar value between 0 and 1. For a unit i, it is defined as the probability of receiving the treatment given the observed covariates:

$$e_i = P(T_i = 1|X_i) \tag{2.8}$$

In practice, however, true propensity scores are rarely known outside of randomized experiments. Propensity scores therefore must be estimated, which is typically done by using logistic regression (Stuart, 2010). Let $D_{i,j}$ be the distance between units i and j. Then, linear propensity score distance is defined:

$$D_{i,j} = |logit(e_i) - logit(e_j)| \tag{2.9}$$

where e_k is the propensity score for unit e_k (Stuart, 2010). Together with the linear propensity score distance measure, it is common to use a propensity score caliper calculated from some standard deviations of the estimated distribution of propensity scores. Rosenbaum and Rubin (1985) generally recommends 0.25 standard deviations, but this value needs fine-tuning and depends on the difference in variance between the treatment and control groups (Stuart, 2010). If the variance in the control group is a lot smaller than that of the treatment group, less standard deviations are necessary. For instance, Mozer et al. (2019) used 0.1 standard deviations of the distribution of propensity score distances in their investigation of the causal impact of bedside transthoracic echocardiography (TTE) on the survival rate of sepsis patients.

2.3.2 Matching Procedure

Once a distance metric has been selected, the next step is to use that distance metric in a matching procedure to match treatment and control units. Stuart (2010) describe five types of procedures for matching in general: nearest neighbor matching, sub-classification, full matching, and weighting. In this study, we are concerned with full matching (Hansen, 2004; Rosenbaum, 1991; Stuart & Green, 2008). Full matching automatically create matched sets of treatment and control units which are similar, where each set contains at least one treated unit and at least one control unit. It does this in an optimal way by minimizing the average distances between each treated unit and control unit within each matched set. Treatment and control units that do not have an adequate enough match (as defined by calipers) are discarded. Full matching allows one to reuse control units for many treatment units and to specify ratios that determine how many control units are able to match with a treatment unit. In this study, we use one-to-one full matching where each treatment unit is matched to a single control unit. Stuart (2010) recommends full matching when estimating the ATT and the number of treatment units are less than the number of control units, which is the case in our study. From now on, following Mozer et al. (2019), we refer to this as optimal one-to-one matching.

2.3.3 Match Quality Analysis

Once a matched sample has been generated by the matching procedure, match quality is evaluated. As explained previously, treatment and control units are matched based on the covariates. Two units are a perfect match if their corresponding covariates are exactly the same. In other words, if the difference between corresponding covariates equals 0. The closer the differences of covariates are to 0, the better the match. As evaluating the match quality of each individual match would be unfeasible, we instead evaluate match quality of the matched sample. The aim is for the empirical distributions of the full set of covariates in the matched sample to be as similar as possible (see Equation 2.7). Note, if these distributions are equal, we have satisfied the ignorability assumption (Stuart, 2010). We refer to this similarity as covariate balance. Covariate balance is typically evaluated by calculating the standardized mean difference of each covariate (Markoulidakis et al., 2020; Mozer et al., 2019; Stuart, 2010). When estimating the ATT, the standardized mean difference (SMD) is defined in the following way for each covariate.

$$SMD_{ATT} = \frac{\overline{X}_{treatment} - \overline{X}_{control}}{sd_{treatment}} \tag{2.10}$$

The numerator is the mean of the treated group minus the mean of the control group, and the denominator is a measure of spread calculated as the standard deviation of the treated group. Note, it is the averages of the treated and control groups *after* matching, where units in the two groups potentially have been pruned. SMD values range from -1 to 1, where a value close to 0 indicates a higher similarity between the two mean values, and therefore a better balance of a covariate. Values between -0.1 and 0.1 are generally regarded as well-balanced (Markoulidakis et al., 2020; Mozer et al., 2019). Covariate balance should be evaluated after each matching procedure and matching procedures that result in highly imbalanced samples should be rejected (Stuart, 2010). If a matching procedure is rejected, step 1-2 should be repeated until a sample that is well-balanced is attained.

2.3.4 Outcome Analysis

Whereas the first three steps represent the design phase, the fourth step is the analysis phase. After the matching in step three has created a matched sample with adequate balance between treatment and control groups, analysis of the outcome and estimation of treatment is done (Stuart, 2010). One could perform regression adjustments using the matched sample in order to "fix" small residual covariate imbalance between the treatment and control groups (Stuart, 2010). One could also do significance tests of the estimated causal effects and perform robustness checks to assess the plausibility of assumptions made when conducting the research. For instance, assessing whether the overlap or ignorability assumptions have been violated. The ignorability assumption can never be tested directly. Researches have therefore developed sensitivity analysis to assess its plausibility and how violations of the ignorability assumption may affect the conclusions of a study (Stuart, 2010). An example of such sensitivity analysis is to estimate the effect on a variable that is known to be unrelated to the treatment (Imbens, 2004). The ignorability assumption is then deemed less plausible if the test indicates that the effect is not equal to zero. There is also no formal way to assess the overlap assumption. Markoulidakis et al. (2020) recommends doing simple checks using summary statistics to assess obvious areas of the covariate distribution where there is a lack of overlap. For instance, comparing the minimum and maximum of the same covariate in the treatment and control groups, or by plotting the distributions on top of each other. If there still exists a lack of overlap after matching, one could remove those units for which no adequate match is found and continue with the estimation of causal effects. In this study, as our contribution is methodology focused, we estimate the ATT (see Equation 2.4) without any further analysis or regression adjustments.

2.4 Text Matching

So far, traditional matching on numerical covariates has been explained using the four steps outlined by Stuart (2010). In this section, we explain how to incorporate text by building on this framework (Mozer et al., 2019). In order to match on text, one needs a representation of text that expresses the corpus in a structured and quantitative form. Additionally, one also needs to choose a distance metric that quantifies how similar two documents are. Step one of text matching therefore specifies a text representation that defines explicitly the features that will be considered covariates, as well as a distance metric to measure the similarity between two documents. This distance metric should ideally focus its attention on the aspects of the text considered most important to account for (i.e. the biggest potential confounders). Examples of text representations include:

- Term-Document Matrix (TDM),
- Statistical Topic Models (STM), and
- Word2Vec.

These different text representations come with different trade-offs between retaining information and interpretability (Mozer et al., 2019). TDMs favor retaining more information at the cost of high dimensionality. Documents that are matched based on their term-vectors will generally be similar with regards to both usage of keywords and topical content. STMs match documents based on topic proportions and thus favor dimension reduction at the potential cost of information loss. The neural network architecture Word2Vec (Mikolov et al., 2013) falls somewhere in between (Mozer et al., 2019).

The choice of distance metric depends on this trade-off between retaining information and interpretability. Exact, coarsed exact and Mahalanobis distance tend to not work well with high dimensional text representations (Mozer et al., 2019; Stuart, 2010), while cosine distance has shown to perform well in combination with TDM-based text representations, which are high dimensional (Mozer et al., 2019). Ideally, the selected representation and distance metric will mimic, or even surpass human judgement of how similar documents are. However, choosing such a text representation and distance metric combination is challenging, due to the aforementioned trade-off between retaining information and interpretability, and the best representation might differ from domain to domain (Mozer et al., 2019).

Once a text representation and distance metric have been selected, text matching practically follows the same procedure as traditional matching, except for the match quality analysis. Due to text being high dimensional, numerical diagnostics can be difficult to interpret when evaluating match quality. For instance, when using a TDM to represent text, each term in the TDM is a covariate. It would be unfeasible to measure the covariate balance on each term, especially as some terms are likely to be more confounding (and thus deemed more important) than others. Hence, because text is highly interpretable, Mozer et al. suggests that one should utilize qualitative evaluation of match quality using human intuition.

Evaluating the match quality of matched documents can be done in different ways depending on your application. In an experimental study, Mozer et al. utilized online crowd-sourcing platforms such as Amazon's Mechanical Turk (MTurk) and the Digital Laboratory for the Social Sciences (DLABSS) to evaluate how text representations and distance metrics correspond to human judgement. In that study, 505 respondents were given a series of 11 paired newspaper articles, including an attention check and an anchoring question, and asked to assign a similarity rating. These respondents were first informed about the nature of the task and then given training on how to evaluate similarity of two documents. Match quality was defined on a scale of 0 (lowest quality) to 10 (highest quality). Final similarity between two documents was then calculated as the average rating that respondents gave. In an applied study that investigated political bias of news media organizations, human workers were tasked to read and evaluate the ideological position of each article on a 5-point Likert scale.

In the two studies above, one could rely on non-expert human coders. However, some applications require expert domain knowledge in order to evaluate whether two documents are similar. In a second applied study, Mozer et al. used text matching when analyzing the effects of bedside transthoracic echocardiography (TTE), a tool used to create pictures of the heart, on the survival rate of sepsis patients. In addition to numerical data on patients, textual data was collected in the form of intake notes and matched on. These intake notes cannot be reliably evaluated by non-expert human coders. Mozer et al. therefore adopted an information retrieval approach where medical professionals mapped the intake notes to a set of clinically meaningful concepts and prognostic factors that could be used to characterize intensive care unit patients. Jaccard similarities over this mapping between matched pairs of documents were then calculated and treated as the gold standard for measuring match quality of matched documents. Furthermore, covariate balance was also analyzed on both quantitative and text-based covariates that medical experts deemed potentially confounding. Each of the text-based covariate was calculated using summary measures based on word counts from the patient-level text documents.

In their application that analyzes the effects of TTE on sepsis patients, Mozer et al. applied two different matching approaches. One that matches patients on only the numerical data, and one that matches patients using both the numerical and the text data. Both approaches used the optimal one-to-one matching procedure. The first approach matched numerical covariates based on estimated propensity scores (PSM) and used a caliper equal to 0.1 standard deviations of the estimated propensity score distribution. The second approach matches intake notes within the same propensity score caliper used for the first approach. By doing this, the space of possible treated-control pairings is first reduced in a way that ensures adequate balance on the numerical covariates. Within this space, by doing text matching one is also able to capture variables that were not recorded numerically, but can be estimated by summary measures of the text (Mozer et al., 2019).

Using TDM-based text representations in combination with the cosine distance metric, Mozer et al. found that text matching within propensity score calipers significantly improved the covariate balance of the text-based covariates, and had similar covariate balance on the numerical covariates, compared to the first approach. Moreover, they found that the effective sample size was larger in the second approach. They attributed this to text-based distances offering a more refined measure of pairwise similarity compared to distances based on propensity scores, when within propensity score calipers. Furthermore, in all of their studies, cosine distance and TDM-based representations produced high quality results and were robust to tuning of parameters, such as the choice of weighting scheme and the degree of bounding applied to TDMs.

3
Methodology

The methodology chapter describes this study's methodological approach. First, we elaborate on the nature of the IMF policy document data and the numerical country-level data. Second, we explain the pre-processing of this data and how observations are put into treatment and control groups. Third, we explain our confounder parsing procedure, which can be divided into two stages. Stage one converts policy documents in PDF format to raw text. Stage two removes treatment text from these policy documents to only end up with confounding text. Thereafter, the causal inference pipeline is explained. This is divided into three sections describing: our matching approaches, how match quality is evaluated and how causal impacts are estimated. Finally, core limitations of the methodology are explained and we finish with an ethical and risk analysis.

3.1 Data Collection

Although our research questions are of a methodological nature, we need a domain area to apply our research to and to extract numerical and textual data from. We have chosen IMF policy evaluation as our area to perform matching on and estimate causal effects. However, since we do not have a thorough understanding of the IMF domain, we use previous research as a guideline to determine a suitable set of numerical outcome- and confounding variables. We build upon Daoud and Reinsberg (2019), who measured the impact of labor issue policies on health using three outcome variables as a proxy for health, which can be seen in Table 3.1 below. For this research, Daoud and Reinsberg (2019) used a bundle of the most impactful covariates that are also confounding variables of the outcomes. Hence, these covariates are impactful in two ways. First, they strongly influence the likelihood of a country engaging in an IMF program, including labor policy conditions. Second, such covariates heavily influence the three health outcome variables. We use 14 of those chosen numerical covariates, which can also be observed in Table 3.1 below.

Table 3.1: Country variables and data sources (Daoud & Reinsberg, 2019)

Variable	Definition and sources
Outcome (t+1, t+3)	
Child mortality	Under-five child mortality (World Bank 2015)
Vaccination index	Index of vaccination, computed as the average vaccination (as percentage of the population) against measles, polio, and diphtheria (World Bank 2015)
Health expenditure	Public-health expenditure as a percentage of total government expenditure (World Bank 2015)
Covariate (t-1)	
Civil war	Incidence of civil war according to UCDP/PRIO definition (Teorell et al., 2020)
GDP per capita (log)	GDP per capita in constant 2005 USD (World Bank 2015)
Population density (log)	Population density, computed as people per sq. km of land area (World Bank 2015)
UNNA population (log)	Population measured by the United Nations (National Accounts Main Aggregates Database)
UNGA vote alignment	Vote alignment of a country with the G7 in the UN General Assembly (Bailey, Strezhnev and Voeten 2015)
Reserves	Reserves in months of imports (World Bank 2015)
Dependency ratio	Dependency ratio, computed as the combined share of the population under age of 14 and above age of 65 in the total population (World Bank 2015)
Freedom House Index	Combined civil liberties and political rights from Freedom House and inverted in scale (higher values are better) (Teorell et al., 2020)
Oil per capita (log)	Oil per capita, computed as oil production in metric tons divided by population (National Accounts Main Aggregates Database, World Bank 2015)
Executive elections	Incidence of executive elections—Database of Political Institutions (Teorell et al., 2020)
Trade (log)	Export minus import as a percentage of GDP (World Bank 2015)
GDP growth	GDP growth in percent (World Bank 2015)
Urbanization	Urban population as a percentage of total population (World Bank 2015)
ODA per capita (log)	ODA per capita in constant 2011 USD (World Bank 2015)

The unit of observation for our causal inference problem is IMF arrangement, which can also be seen as a *country-year* pair, where *year* is the year when the first condition of an arrangement was implemented, or start year. We define $t = 0$ as the start year of an arrangement. Following Daoud and Reinsberg (2019), we lag the covariates by one year ($t - 1$) and measure outcomes one ($t + 1$) and three years ($t+3$) after the start year. These lags are put in place to capture the lagged effects of large economic, political and societal changes, which often take some time to unfold and show up in the data.

We use IMF policy documents as our textual data and have access to a corresponding hand-annotated dataset, provided by a research team at the University of Cambridge. The aim of this dataset is to allow systematic examination of IMF-mandated adjustment policies for each country in every year. It is a disaggregated dataset of all conditions included in IMF programs per calendar year, and it provides information on their implementation. The dataset concerns policy papers ranging from from 1980 to 2014, which consists of 744 different IMF arrangements in 138 countries. Each row in the dataset describes a condition associated with an arrangement, and each arrangement contains one or more conditions a country must implement. In total, the dataset consists of 32,260 rows and 20 columns. We are only interested in a subset of these, presented in Table 3.2.

Table 3.2: IMF dataset variables of interest

Variable	Description
Arrangement ID	Identifier of arrangement.
Country Code	World Bank 3-letter country code.
Text	Condition text (typically one sentence), copied and pasted from the PDF document where the condition appears. We refer to this as treatment text (or t_text).
Year	Year when the condition was implemented.
Type	Type of condition. A condition is classified as one out of five different values. These types of conditions can further be broken down into quantitative and structural conditions.
Waiver	Number of waivers granted to a condition.
Policy Area	Policy area of condition as coded by the researchers. In total there are 13 policy areas, ranging from social policy to external debt issues.
Condition Source Document	Refers to the PDF where the condition appears.

3.2 Data Pre-processing

Consider Table 3.2. We use Arrangement ID to group rows with conditions associated with the same arrangement. All arrangements that are ongoing, or arrangements with one or more waivers, are assumed not to have implemented all of its conditions yet. These arrangements are therefore removed. Furthermore, we limit our research by only focusing on structural conditions and not quantitative conditions. Conditions of type Quantitative Performance Criterion (QPC) and Indicative Benchmarks (IB) are therefore removed, while conditions of type Prior Action (PA), Structural Performance Criterion (SPC) and Structural Benchmark (SB) are kept. Keeping conditions of type QPC and IB would make the process of removing treatment text from policy documents unfeasible, as treatment text (Text) for these conditions have been standardized in the dataset and cannot be found in the PDFs. This pre-processing is done in Python and results in 522 arrangements and 14847 conditions.

An arrangement can be associated with multiple policy documents in the form of PDFs. Ideally, we would remove all treatment texts from these policy documents and merge their confounding texts into one text. However, it is possible that policy documents of an arrangement are updated versions of each other, but the Text field in the dataset only points to the policy document where the condition first appears. We therefore cannot tell with certainty if all treatment texts actually are removed after merging confounding texts from these individual policy documents. For this reason, we will parse and remove treatment texts from all PDFs of the 522 arrangements for the sake of evaluating how well we are able to parse PDFs, but during matching we only use one policy document from each arrangement. For matching, PDFs with conditions of policy area labor issue (LAB) are prioritized. If there are multiple PDFs with conditions of this policy area, the earliest of those PDFs associated with the arrangement is chosen. If an arrangement does not have conditions of policy area LAB, the earliest PDF associated with the arrangement is chosen. Arrangements with one or more conditions of policy area LAB are put in the treatment group. Remaining arrangements are put in the control group. This results in 271 treatment units and 251 control units.

Because the country-level data dates back to 1980, it is prone to have a lot of missing values. The number of missing values for each country-level variable is shown in Table 3.3 below. Health expenditure and reserves are two variables that stand out with more than 100 missing values. Removing each row from the dataset that contains a cell with a missing value would remove 271 rows, which is more than half of our dataset. The remaining 245 rows would constitute too little data to be of use. Instead, we impute the missing values using the R package Amelia II (Honaker et al., 2019a). Amelia II uses a bootstrap and expectation maximization (EM) algorithm to impute missing values from a dataset and outputs multiple imputed datasets. In our case, we tell Amelia II to output ten different datasets with imputed values. We then run our causal inference pipeline on each individual imputed dataset and combine the results. This is explained subsequently.

Table 3.3: Missing values of country variables

Variable	Missing values
Child mortality (t+1)	16
Child mortality (t+3)	21
Measles (t+1)	10
Measles (t+3)	18
Polio (t+1)	18
Polio (t+3)	26
Diphtheria (t+1)	18
Diphtheria (t+3)	26
Health expenditure(t+1)	183
Health expenditure(t+3)	149
Civil war	9
GDP per capita	39
Population density	12
UNNA population	25
UNGA Vote Alignment	23
Reserves	116
Dependency ratio	18
Freedom House Index	35
Oil per capita	36
Executive elections	34
Trade	38
GDP growth	50
Urbanization	14
ODA per capita	7

3.3 Confounder Parsing

IMF policy documents consist of both treatment text and confounding text. We only have access to these policy documents as PDFs. In the text matching methods we only want to adjust for the confounding text in policy documents — if matching is done on treatment text, we risk being exposed to selection bias, which would make our estimations after matching less credible. The purpose of confounder parsing is therefore two-fold:

1. Convert policy documents in PDF format to raw texts in order to be able to use the texts in our matching methods.
2. Remove treatment texts from policy documents to only end up with confounding text to match on.

To what degree we are able to remove treatment texts from policy documents depends on how well we are able to convert PDFs to raw text. It turns out this is not a trivial task. Therefore, we investigate different parsing strategies. These parsing strategies only differ in how they remove treatment texts from policy documents.

Parsing is therefore done in two stages and both stages are implemented in Python.

The first stage involves converting policy documents in PDF format to raw text documents and mapping each treatment text associated with a policy document to a page index and section index (or page-section index). The second stage involves removing treatment texts from policy documents to end up with only confounding text. Breaking down the confounder parsing procedure into two stages serves multiple purposes. First, we get descriptive statistics on how many treatment texts we were able to map to page-section indices. This way we can evaluate how well the mapping performs. This is also useful for parsing strategies that do not rely on the aforementioned mapping in order to remove treatment texts from policy documents, as it is an indication for how well the PDF-to-text conversion performed. No mappings for a policy document might indicate that the tool used to convert the PDF to raw text did not do a good job (perhaps because of the way the PDF was generated). Second, since the first stage is done the same way for each parsing strategy, we save computational time; converting PDFs to raw text can be a slow procedure, especially if PDFs are not searchable and need optical character recognition (OCR'ing) in order to extract text.

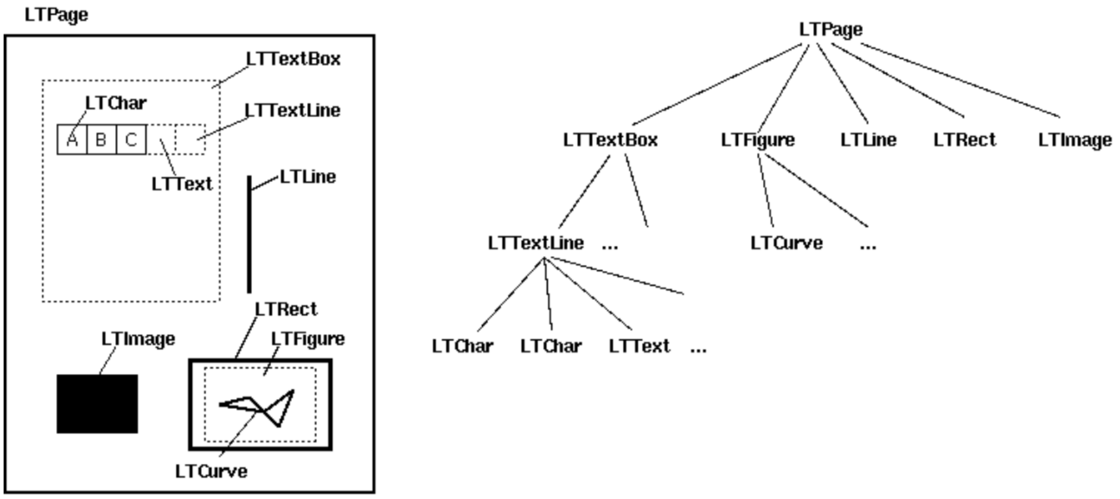
3.3.1 Stage One

Stage one is done in 4 steps:

1. Convert policy document in PDF format to raw text.
2. Clean raw text.
3. Clean treatment texts associated with policy document.
4. Map treatment texts to page-section indices.

Step 1. When converting a policy document in PDF format to raw text, we want our conversion tool to output a stream of meaningful text pieces, such as paragraphs and sentences, page by page. When it comes to text, however, a PDF file is only aware of its characters and their placement, represented as a sequence of bytes (Adobe, 2006). Characters that make up a paragraph are no different from those that make up a description of a figure, page footer or a table. For this reason, extracting meaningful pieces of text from a PDF is not trivial. One tool that attempts to reconstruct some of these meaningful pieces of text is pdfminer.six (Guglielmetti, 2020). It uses a layout analysis algorithm that groups characters into words and lines, lines into boxes and finally groups those text boxes hierarchically, page by page. See Figure 3.1 below, taken from pdfminer.six's user manual (Shinyama et al., 2019). From now on we refer to these text boxes as *sections* in a policy document.

Figure 3.1: Pdfminer.six's layout analysis algorithm outputs a hierarchy of layout objects



A section is defined by the space around text and texts that are close to each other are grouped together in the same section. We then have the ability to read the raw text of a PDF, section by section, instead of a long stream of characters. Being able to do this out of the box provides many benefits. Text pieces that are close to each other also tend to belong to each other semantically. Hence, we are able to read paragraphs without being concerned with the original policy document structuring its text into multiple columns on a page, because each column will be contained within its own section.

To our knowledge, other PDF-to-text conversion tools than pdfminer.six, such as pdftotext (Palmer, 2021) and pdfreader (Polshcha, 2021), do not support this behaviour. Instead, you read the text as a single stream and have to define what a section is yourself. For these tools, single column pages are relatively easy to parse, as each paragraph is usually separated by two consecutive new-line characters ('\n'). However, parsing pages with multiple columns is difficult because of the inconsistency of how text is extracted from PDFs when our policy documents range from 1980 to 2014. The effect of not handling multi-column pages would mean we are merging columns on a row by row basis. Consequently, word order on these pages is lost, which negatively affects step 4 when mapping treatment texts to page-section indices, as we cannot rely on string comparisons. It turns out that most treatment texts for structural conditions appear in tables, which contain multiple columns. Handling columns is therefore not something we should neglect.

For all the reasons laid out so far, pdfminer.six is the preferred choice of tool for converting PDF to text. However, pdfminer.six also has its cons. A pilot study of ours showed that it did not interpret text in tables well when the font size was smaller compared to the rest of the text. It also does not interpret older PDFs well. We suspect this is related with the quality of the tool generating PDFs. Furthermore, because pdfminer.six looks at the source code of a PDF, it does not know how to interpret PDFs that have been scanned. We assume that all PDFs for which step 4 cannot find any treatment to section mappings are scanned. For these PDFs, we will repeat step 1-4, but using the tesseract-ocr tool (Smith, 2021) for extracting text in step 1. Before using this tool, all pages of a PDF must first be converted to images. The tesseract-ocr tool then extracts text from these images using an OCR engine based on LSTM neural networks. We used the tessdata-best training data for this engine — the most accurate training data as of September 15, 2017. For this iteration of stage 1, we do not handle multi-column pages and define a section as text separated by two consecutive new-line characters. Not handling multi-column pages is a limitation.

Step 2. Once a policy document has been converted from PDF format to raw text, we clean each section in the raw text from unwanted tokens. For each section, we remove all tokens that are not a letter in the English alphabet or white space. Lastly, we make all text lowercase and remove all unnecessary white space (e.g. trailing white space or multiple space tokens between words). A cleaned section ends up being a space-separated list of tokens only containing letters from 'a' to 'z'. Tokenization is done using the Python package spaCy (Explosion, 2021).

Step 3. We use the same cleaning procedure on the treatment texts as on the sections in the raw policy document text. Cleaning both treatment texts and sections in the policy documents this way, we relax the constraints for what a mapping is in step 4.

Step 4. The last step of stage one is to map treatment texts to page-section indices. For each policy document text, this is done in the following way. We read the policy document page by page and section by section. For each section, we check if any of the treatment texts associated with the policy document is a valid mapping with the section according to mapping levels described in Table 3.4 below. A section is referred to as s_text and a treatment text is referred to as t_text.

Table 3.4: Valid treatment-to-section mappings and their levels

Name	Level	Description
EXACT	14	t_text equals s_text.
EMBED	13	t_text is a sub-string of s_text.
WORD_PERC_95	12	95% of words in t_text appears in s_text.
LEMMA_PERC_95	11	95% of lemmas in t_text appears in s_text.
WORD_PERC_90	10	90% of words in t_text appears in s_text.
LEMMA_PERC_90	9	90% of lemmas in t_text appears in s_text.
WORD_PERC_85	8	85% of words in t_text appears in s_text.
LEMMA_PERC_85	7	85% of lemmas in t_text appears in s_text.
WORD_PERC_80	6	80% of words in t_text appears in s_text.
LEMMA_PERC_80	5	80% of lemmas in t_text appears in s_text.
WORD_PERC_75	4	75% of words in t_text appears in s_text.
LEMMA_PERC_75	3	75% of lemmas in t_text appears in s_text.
WORD_PERC_70	2	70% of words in t_text appears in s_text.
LEMMA_PERC_70	1	70% of lemmas in t_text appears in s_text.

If a treatment text has a valid mapping with a section according to Table 3.4, we associate this treatment text with corresponding page index, section index and mapping level. EXACT and EMBED are string comparisons, whereas the remaining mappings calculate the percentage of individual words/lemmas in the treatment text that appears in a section, regardless of the order of these words/lemmas. Tokenization/lemmatization of sections into words/lemmas is done using the Python package spaCy (Explosion, 2021). We anticipate that a lower mapping level has a higher risk of being a false positive than higher mapping levels. For example, a mapping of level 14 is guaranteed not to be a false positive since the treatment text and section are exactly the same, character by character. On the other hand, a mapping level of 1 is likely to be a false positive if the section consists of many sentences and the treatment text only of a few words. For this reason, a mapping with a higher level is always preferred. Furthermore, a treatment text can only be mapped to at most one section. Since we parse policy documents top-down, there is a possibility that we find a better mapping for a treatment text at a later state. We therefore allow updates of a mapping if and only if the level of the new mapping is greater than that of the previous one.

3.3.2 Stage Two

The goal of stage two is to remove all treatment text from the policy documents to end up with only confounding text. We have four parsing strategies with different trade-offs for this:

- section parsing (SP),
- page parsing (PP),
- count subtraction parsing (CSP), and
- full subtraction parsing (FSP).

SP reads each policy document page by page and section by section. Denote the current page as p_i and the current section as s_i. For each policy document, whenever we read a section s_i that is mapped to a treatment text associated with the policy document, we remove that section. This is done using the page-section indices that we kept track of in stage one. PP is done in the same way, but instead of just removing section s_i, we remove page p_i as well as pages p_{i-1} and p_{i+1}. The reasoning for removing pages p_{i-1} and p_{i+1} is because treatment texts tend to appear close to each other in IMF policy documents. This is especially true for treatment texts appearing in tables; if a treatment text appears in a table, there is a good chance more treatment texts appear in the same table. Furthermore, tables can span multiple pages. Hence, if we are able to find a mapping for treatment text a, but not for treatment text b, it is likely we would still remove b from the policy document, although it is not shown in the statistics from stage one.

Since we keep track of the level of each mapping, we can decide to only remove sections or pages for which the mapping level is above, or equals, a certain threshold. For example, setting this threshold to 10 for SP, would only remove sections from a policy document that have a mapping level between 10 and 14. Sections with no mapping or mapping level below 10 would still be in the policy documents. Setting this thresholds to 0 effectively means that we leave all policy documents as is. In this study we limit ourselves to the thresholds 1, 5 and 10, for both SP and PP.

CSP and FSP do not remove text from policy documents on a per section or page level. Instead, they look only at the treatment texts and remove individual words from policy documents. In CSP, for each policy document, we create a vector of words from all treatment texts associated with the policy document. The individual words of this vector are then removed from the policy document. For example, if the word 'government' appears three times in the treatment vector, we would remove three instances of 'government' in the policy document vector. FSP is similar to CSP, but instead of creating a vector of the words from all treatment texts associated with the policy document, we create a set of all unique words from the treatment texts. We then remove *all* instances of words in the treatment set from the policy document. For example, if the word 'government' appears in the treatment set, we would remove all instances of 'government' in the policy document vector.

As a last step of stage two, we clean the confounding text even more by removing stop words using spaCy (Explosion, 2021) and words that represent Roman letters/numerals. This serves as a pre-processing step for text matching. We do not want texts to be seen as similar based on words without semantic meaning, and removing stop words has long been seen as a best practice.

In total, this study investigates how covariate balance after matching is affected using 8 different parsing methods: CSP, FSP and SP and PP with mapping level thresholds 1, 5 and 10. All of these different parsing methods have different tradeoffs. While SP favors keeping more of the confounding text in policy documents, it is prone to also keep more treatment text. The reverse is true for PP. PP is more likely to remove more treatment text, but at the expense of also removing a lot of confounding text. Furthermore, the lower mapping level threshold you set for these two parsing methods, the more treatment and confounding text is removed from the policy documents.

FSP and CSP are the most precise parsing methods in that they remove all treatment texts and remove little to no confounding text. However, this assumes that all of the treatment texts have been captured in the dataset and that no latent treatment text exists in the policy documents. Furthermore, it also sets limitations on what kind of text representation one can use for text matching. CSP and FSP remove treatment words arbitrarily from policy documents. Hence, it requires text representations for matching to not care about word order, such as Bag-of-Words (BoW) models. In that sense, SP and PP are more general and would work for all kinds of text representations.

3.4 Causal Inference Pipeline

3.4.1 Matching

We apply three different matching approaches to our data: one that matches arrangements only on numerical covariates and ignore text covariates (baseline), one that matches arrangements only on text covariates and ignore numerical covariates, and one that matches arrangements using both the numerical and text covariates. In the first approach, following Mozer et al. (2019), we match treated and control units using optimal one-to-one matching (Hansen & Klopfer, 2006) on estimated propensity scores of the covariates in Table 3.1. We enforce two calipers. The first caliper ensures that treated and control units only match if their arrangements' start year are within three years of each other. Ideally, we would like to match treated and control units that have been observed at the same point in time. However, a pilot study of ours showed that this limits the number of possible matches to a degree where match quality becomes too poor in order to make any credible causal effect estimates. By relaxing this caliper to three years, a trade-off is made to increase the number of possible matches. Following Mozer et al. (2019), we also enforce a propensity score caliper. This caliper is equal to 0.3 standard deviations of the estimated distribution of propensity scores, which discards any treated units for whom the nearest control unit is not within a suitable distance. Mozer et al. (2019) used 0.1 standard deviations, but 0.3 standard deviations worked better for our data. Furthermore, we make sure that treated and control units of the same country do not match. This is done by building a distance specification where treated units are infinitely far away from control units if they share the same country code.

In the second matching approach, we perform optimal one-to-one text matching using the same constraints on start year of arrangement and country code, but we only match treated and control units based on distances between policy documents. The third approach is similar to the second approach, but text matching is done within the propensity score calipers, following Mozer et al. (2019). From now on, we refer to these matching approaches as

1. propensity score matching (PSM),
2. text matching without propensity score calipers, and
3. text matching within propensity score calipers.

In order to do text matching, we need to specify a text representation and a corresponding distance metric for the second and third matching approaches. Because of time constraints, we limit ourselves to trying one text representation and one distance metric. Mozer et al. (2019) found that cosine distance and TDM-based representations produced high quality results in all of their applied examples. Furthermore, since TDM is a BoW model, it lends itself to the count subtraction parsing (CSP) and full subtraction parsing (FSP) strategies explained previously. Based on this, this study use cosine distance and TDM text representation with six different combinations of bounding and weighting schemes. See Table 3.5 below.

Table 3.5: TDM settings used in text matching

Name	Bounding scheme	Weighting scheme
tdm1	(4, 1000)	Term frequency
tdm2	(4, 1000)	Term frequency - inverse document frequency
tdm3	(4, 10)	Term frequency - inverse document frequency
tdm4	(10, 500)	Term frequency - inverse document frequency
tdm5	(0, Inf)	Term frequency
tdm6	(0, Inf)	Term frequency - inverse document frequency

The bounding scheme determines the subset of the vocabulary that will be included. Each TDM setting has a lower and an upper bound. Terms that appear in less documents than the lower bound or in more documents than the upper bound are discarded. By setting bounds we have the ability to eliminate extremely rare and/or extremely common terms in our vocabulary. Setting the lower bound to 0 and the upper bound to infinity, as done for tdm5 and tdm6, is effectively the same as using the original vocabulary, or using no bounding scheme. The weighting scheme determines the numerical rule for how the values of the text covariates are measured. Following Mozer et al. (2019), we consider standard term-frequency (TF) weighting and term frequency - inverse document frequency (TF-IDF) weighting.

For each of the eight confounder parsing methods, we try all six TDM settings for both of the text matching approaches. This results in $8 * 6 * 2 = 96$ different text matching methods. Together with the first matching approach, we have a total of 97 matching methods. We generate matches for all these matching methods on all ten imputed datasets. When doing match quality analysis and estimating causal effects, the results of the same matching methods from each imputed dataset are then combined into one result. How the imputed datasets are combined is explained subsequently. The code for running all matching methods is written in R and is heavily inspired by Mozer et al. (2019) using the optmatch package (Hansen et al., 2019).

3.4.2 Match Quality Analysis

After matching, match quality is analyzed. Following Mozer et al. (2019), match quality is analyzed by looking at the covariate balance for each covariate that is a confounder of our outcomes. We limit ourselves to only analyze the covariate balance on numerical covariates, found in Table 3.1, for all matching methods. Although it is possible to also analyze match quality of text covariates, such as key terms known to be confounding by experts (Mozer et al., 2019), we lack the domain knowledge needed to decide what such text covariates might be for our IMF application. We also lack the time to and resources to do any qualitative evaluation of match quality when matching on text.

Standardized mean difference (SMD) is used to measure covariate balance for each covariate (see Equation 2.10). We use the cobalt package in R (Greifer, 2021) to calculate the SMD of all covariates simultaneously. Following Mozer et al. (2019), standard error (SE) for each SMD is also calculated in order to get an idea of how precise our estimates are of the mean differences.

$$SE = \sqrt{\frac{sd^2_{treatment}}{n_{treatment}} + \frac{sd^2_{control}}{n_{control}}} \tag{3.1}$$

SE is calculated as the square root of the sum of the treatment variance divided by the number of treatment units and the control variance divided by the number of control units. Because of our small sample size, we expect the standard errors to be large.

For each matching method, SMD and SE are calculated for all covariates in all imputed datasets. SMDs and SEs generated from the same matching method in all imputed datasets are then combined into one set of results. This is done using the mi.meld function provided by the R package Amelia II (Honaker et al., 2019a). The mi.meld function uses the rules in Rubin (1987) for "combining a set of results from multiply imputed datasets to reflect the average result, with standard errors that both average uncertainty across models and account for disagreement in the estimated values across the models" (Honaker et al., 2019b).

3.4.3 Causal Effect Estimations

After matching, we estimate causal effects by calculating the Average Treatment Effect on the Treated (ATT), for each outcome of interest (see Equation 2.4). The purpose of estimating causal effects is not primarily to speculate about the impact of IMF's labor issue conditions on health, but to investigate how the matching methods differ in relation to each other with regards to these point estimations. In order to say something about the actual effects, one would need to do more careful analysis, e.g. significance tests, robustness checks and sensitivity analysis (Stuart, 2010).

At this stage, matching has already been done. Equation 2.4 for estimating the ATT can therefore be rewritten as equation 2.4 — the difference between the averages of the outcomes from the treatment and control groups.

$$ATT = \overline{Y}_{treatment} - \overline{Y}_{control} \tag{3.2}$$

We estimate causal effects for the outcome variables in Table 3.1, measured one (t+1) and three years ($t + 3$) after the start year of an arrangement. A standard error is calculated for each estimate using equation 3.1. Since our sample size is small, we expect the standard errors to be large. Similar to the match quality analysis, we estimate causal effects for all matching methods and all imputed datasets. The results of corresponding matching methods in imputed datasets are then combined into one using the mi.meld function provided by the R package Amelia II (Honaker et al., 2019a).

3.5 Limitations

This section elaborates on the most important limitations that we are aware of. First, because we are using text in our matching methods, we are limited to observations that have text. In our case, since we are using IMF policy documents for text, we are limited to country-year observations that are associated with an IMF program. Consequently, we are forced to estimate the ATT and have access to a significantly smaller dataset. On the other hand, if one would only use numerical data, one would have been able to use country-level data for all possible countries measured every year from 1980 to 2014.

Second, because we have country-year observations, we would preferably want to treat the data as panel data when matching in service of causal inference. Instead, we are treating the data as cross-sectional, but make sure not to match observations of the same country. To our knowledge, packages in R are not mature enough to handle text matching with panel data and we lacked the time to create our own implementation.

Third, Mozer et al. (2019) argues that one should really evaluate match quality on text matching qualitatively by reading and comparing documents. For instance, they utilized online crowd-sourcing platforms such as Amazon's Mechanical Turk (MTurk) and the Digital Laboratory for the Social Sciences (DLABSS) to evaluate how text representations and distance metrics correspond to human judgement. Time constraints limit us to only measure match quality quantitatively by analyzing covariate balance on numerical confounders, identified by existing literature. Furthermore, there might be specific confounding text covariates (e.g. key terms) that are especially important to balance (Mozer et al., 2019). As existing literature has not identified any such text covariates and we lack the domain knowledge, we only analyze covariate balance on numerical confounders.

Moreover, we only use TDM-based text representations in combination with cosine distance metric, while many other combinations exist that could yield different results. One could for example use word embeddings, neural networks or topic models, combined with different kinds of distance metrics. Hence, this study does not uncover whether confounder parsing is robust across different combinations of text representations and distance metrics. Additionally, in practice none of the confounder parsing methods guarantee that all treatment texts are removed. For instance, there might be latent treatment text that our confounder parsing procedure does not capture. Therefore, we are at risk of being exposed to selection bias, although that bias has been reduced drastically nevertheless.

Furthermore, this study is limited to only focus on structural conditions and not quantitative conditions. These types of conditions require more involved parsing techniques, using machine learning to predict if a section should be removed from a policy document. We neither had access to, or time to implement and train, such a machine learning model. Finally, although an IMF arrangement is represented by multiple PDF documents, we are only able to use one document per arrangement. Hence, we are unable to use all available text data that might confound the outcomes. As described in Imbens and Rubin (2015), one should aim to condition on all available data that could contain valuable confounding information when estimating causal effects using observational data.

3.6 Ethics and Risk Analysis

Since we will not deal with personal data and our findings will not explicitly cause harm to people, we don't have any major ethical and risk concerns that need to be taken into account. However, if our methodology is flawed and the findings project the IMF or certain countries in a bad light, this could cause them unjustifiable reputational harm. Incorrect results could also lead other researchers or policy makers to base their analysis on these findings and make unsound decisions that impact the general population.

4

Results

This chapter first present the confounder parsing results. Thereafter, the covariate balance findings are explained. Finally, we examine the causal effects of the IMF public sector conditions.

4.1 Confounder Parsing

We look at how well stage one of our confounder parsing procedure was able to map treatment text to sections for all mapping level thresholds. Remember, if a treatment text has a valid mapping with a section, we associate this treatment text with corresponding mapping level. EXACT (level 14) and EMBED (level 13) are string comparisons, whereas the remaining twelve mappings calculate the percentage of individual words/lemmas in the treatment text that appears in a section, regardless of the order of these words/lemmas. The higher the mapping level, the more strict we anticipate the mapping is. Below is an example of successful mapping between a treatment text and section text at level 13, where the treatment text is a sub-string of the section text:

> **Treatment text:** "Initiate work on the Civil Service Transformation program (payroll survey)"
>
> **Section text:** "Civil service reform Initiate work on the Civil Service Transformation program (payroll survey)"

In contrast to above example, below is an example of a successful mapping between a treatment text and section text at level 1, where 70% of the words in the treatment text exist in the section text:

> **Treatment text:** "Agree with the World Bank and Fund on annual targets for a three-year state-owned enterprise (SOE) reform program for equitization, divestiture, and liquidation, covering 10 percent of SOEs 2019 debt."
>
> **Section text:** "Under the plan, SOEs subject to enterprise-specific reform measures comprise 31 percent of total SOE employment, 11 percent of state capital, and 10 percent of SOE debt. These SOEs are mainly small to medium scale. During the first- year PRGF program, the targets are to

> equitize, divest, or liquidate/close 600-650 enterprises and merge/consolidate 80-90. The cost of safety nets (for approximately 250,000 redundant workers expected over three years, or 13 percent of SOE employment) and debt resolution will be covered by the government 2019s own and external concessional resources (see Box 2). Detailed implementation of this plan will be monitored under the Bank 2019s PRSC."

For all 1605 PDFs parsed, we were able to find mappings between treatment texts and sections for all treatment texts in 1059 of the PDFs, and for at least one treatment text in 1596 of the PDFs, leaving 9 PDFs with no mappings. Table 4.1 below describes treatment level parsing results for all PDFs parsed.

Table 4.1: All treatment level parsing results.

Name	Level	Mappings per level	Cumulative mappings
EXACT	14	4446	4446
EMBED	13	2395	6841
WORD_PERC_95	12	985	7826
LEMMA_PERC_95	11	249	8075
WORD_PERC_90	10	810	8885
LEMMA_PERC_90	9	241	9126
WORD_PERC_85	8	580	9706
LEMMA_PERC_85	7	274	9980
WORD_PERC_80	6	633	10613
LEMMA_PERC_80	5	284	10897
WORD_PERC_75	4	473	11370
LEMMA_PERC_75	3	266	11636
WORD_PERC_70	2	365	12001
LEMMA_PERC_70	1	257	12258
REST	0	2589	14847

For all PDFs, stage one of the confounder parsing procedure was able to find 12258 mappings out of 14847 possible (82.56%). The remaining 2589 (17.44%) treatment texts were not mapped to any sections. Higher mapping levels tend to have a larger number of mappings. 46.07% of treatments texts were mapped to a section using string comparisons (EXACT and EMBED).

In our causal inference problem we limited ourselves to only use one PDF for each arrangement. We therefore also look at how stage one of the confounder parsing procedure performs for this subset of PDFs. For all 522 PDFs used during matching, we were able to find mappings between treatment texts and sections for all treatment texts in 307 of the PDFs, and for at least one treatment text in 519 of the PDFs, leaving 3 PDFs with no mappings. Table 4.2 below describes treatment level parsing results for the PDFs used during matching.

Table 4.2: Treatment level parsing results for PDFs used in matching.

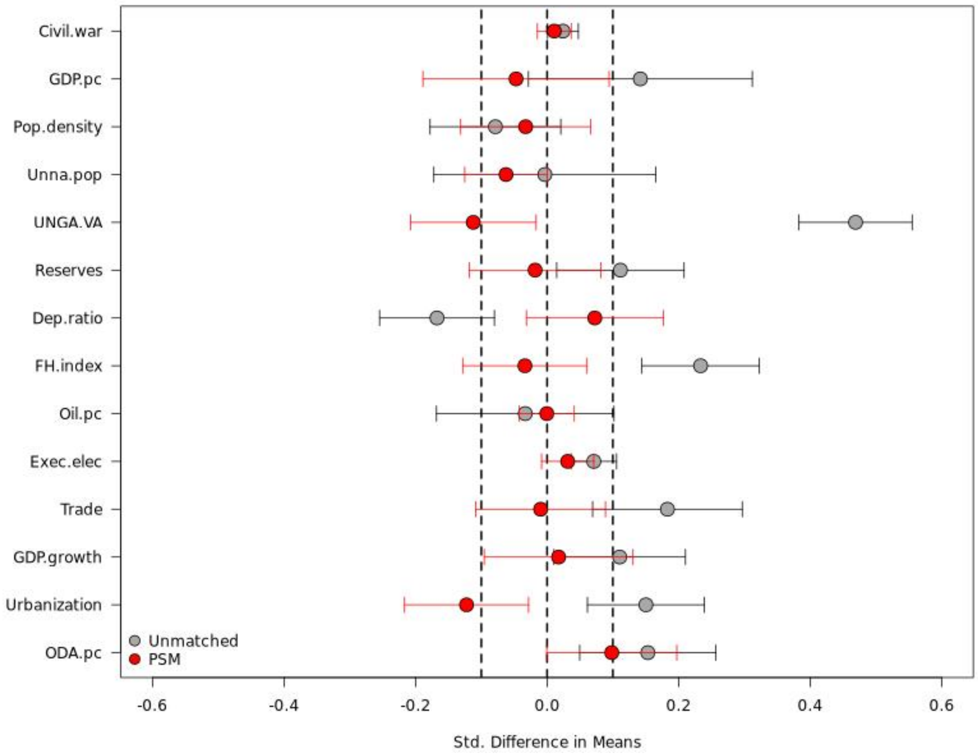
Name	Level	Mappings per level	Cumulative mappings
EXACT	14	1562	1562
EMBED	13	984	2546
WORD_PERC_95	12	381	2927
LEMMA_PERC_95	11	88	3015
WORD_PERC_90	10	321	3336
LEMMA_PERC_90	9	102	3438
WORD_PERC_85	8	248	3686
LEMMA_PERC_85	7	120	3806
WORD_PERC_80	6	265	4071
LEMMA_PERC_80	5	118	4189
WORD_PERC_75	4	173	4362
LEMMA_PERC_75	3	99	4461
WORD_PERC_70	2	141	4602
LEMMA_PERC_70	1	75	4677
REST	0	1125	5802

For the PDFs used during matching, stage one of the confounder parsing procedure was able to find 4677 mappings out of the possible 5802 (80.61%). The remaining 1125 (19.39%) treatment texts were not mapped to any sections. Higher mapping levels tend to have larger numbers of mappings. 44.88% of treatments texts were mapped to a section using string comparisons (EXACT and EMBED). Comparing the treatment level parsing results in Table 4.1 with Table 4.2, it can be observed that they perform roughly the same percentage-wise.

4.2 Covariate Balance

We investigate the covariate balance for all combinations of our three matching approaches. Out of all 97 matching methods, there is only one method matching on only numerical covariates (i.e. the baseline propensity score matching), but 48 combinations for each of the two text matching approaches. Hence, we will only present the covariate balance of a handful of the text matching methods. The covariate balance for the propensity score matching (PSM) is depicted in Figure 4.1 below. Gray dots show covariate balance before matching and red dots show covariate balance after PSM. Values close or equal to 0 is desirable as it indicates the best covariate balance.

Figure 4.1: Covariate balance for propensity score matching



We see an overall improvement on covariate balance after PSM compared to the unmatched covariate balance. After PSM, all but two of the covariates have a standardized difference in means (SMD) in the -0.1 to 0.1 range and are therefore considered well balanced. The UNGA vote alignment (UNGA.VA) and Urbanization covariates are considered unbalanced because they are slightly below the -0.1 border. Oda per capita (ODA.pc) is on the 0.1 border, but still considered balanced. However, the majority of covariates have large standard error bars that reach outside of the -0.1 to 0.1 bound, sometimes even down to -0.23 and up to 0.2. The only covariates with standard error bars within the -0.1 to 0.1 range are Civil war, Executive elections (Exec.elec) and Oil per capita (Oil.pc). Civil war and Executive elections are the only two binary covariates, whereas Oil per capita is continuous. Only one covariate got less balanced after PSM. Population (UNNA.pop) was perfectly balanced before matching and got less balanced after PSM. In the following two figures we look at covariate balance on two parsing-TDM combinations for the approach which does text matching within propensity score calipers.

Figure 4.2: Covariate balance for text matching within propensity score calipers — tdm1 with full subtraction parsing

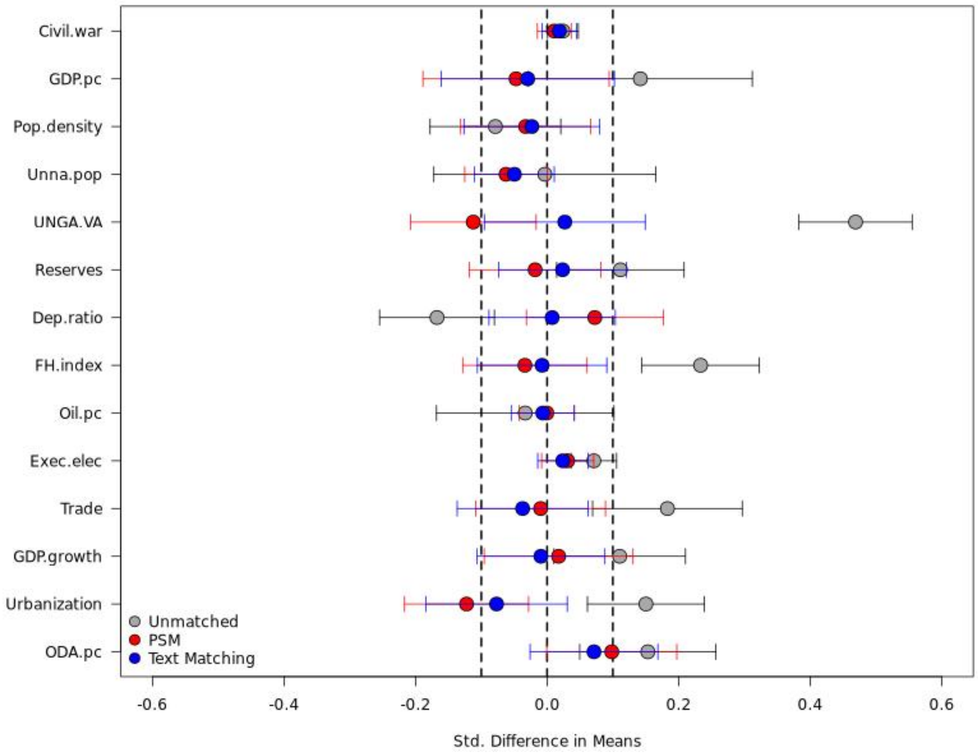


Figure 4.2 above depicts the covariate balance of the combination that use text after full subtraction parsing (FSP) and a TDM with a term frequency (TF) weighting scheme and (4, 1000) bounding scheme. (See tdm1 in Table 3.5.) Blue dots show covariate balance after text matching. Covariate balance of the text matching method in Figure 4.2 is an overall improvement compared to that of the PSM in Figure 4.1. All covariates improved their balance except for Oil per capita (Oil.pc) and Trade. These two covariates are only slightly less balanced. Furthermore, all covariates are considered balanced after the text matching method. However, the standard error bars are still reaching outside of the -0.1 to 0.1 range for most covariates.

Figure 4.3: Covariate balance for text matching within propensity score calipers — tdm6 and page parsing using a threshold of mapping level 1

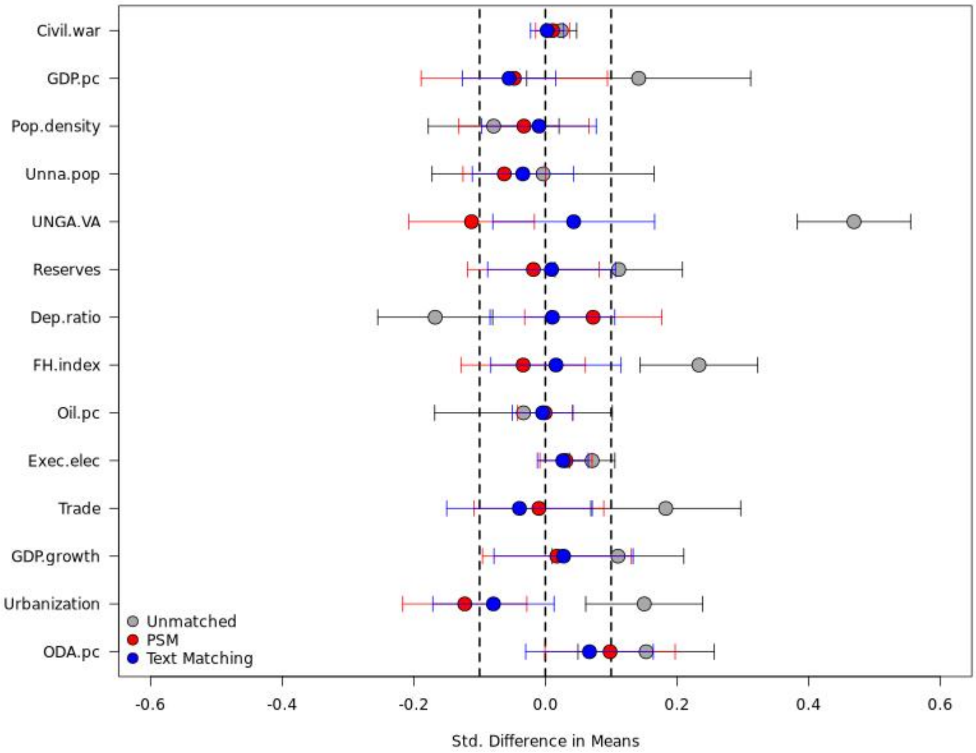


Figure 4.2 above depicts the covariate balance of the combination that uses text after page parsing (PP) with a mapping level threshold of 1 and a TDM with a term frequency - inverse document frequency (TF-IDF) weighting scheme and no bounding scheme. (See tdm6 in Table 3.5 and mapping level 1 in Table 3.4.) Comparing the two different text matching methods in Figure 4.2 and Figure 4.3, we see that the covariate balance is approximately the same for all covariates. This observation is true for all 48 text matching methods that match arrangements within propensity score calipers. For all text matching methods within propensity score calipers, all covariates are within the -0.1 to 0.1 range. Only Urbanization is sometimes placed slightly below the -0.1 bound.

Figure 4.4: Covariate balance for text matching without propensity score calipers — tdm6 and page parsing using a threshold of mapping level 10

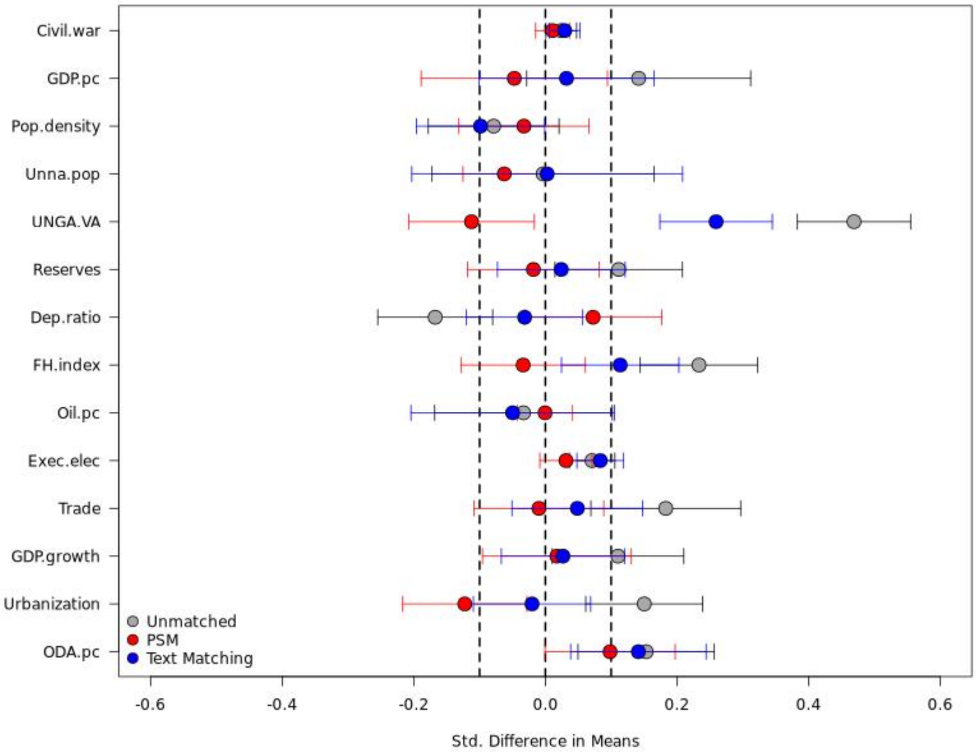


Figure 4.4 above depicts the covariate balance of a text matching method that matches arrangements on only text covariates (i.e. text matching without propensity score calipers). It performs text matching using text after page parsing (PP) with a mapping level threshold of 10 and the same TDM setting as the one in Figure 4.3 (i.e. tdm6 in Table 3.5). This text matching method performs significantly worse compared to the previously presented text matching methods. Only four of the 14 covariates are better balanced compared to corresponding PSM balance. Furthermore, the standard error bars seem to be larger than the text matching methods in Figure 4.2 and Figure 4.3. Moreover, all 48 combinations of text matching without propensity score calipers perform approximately the same with regards to covariate balance.

Table 4.3: Effective sample sizes.

Matching approach	Effective sample size
Before matching	251
PSM	169
Text matching within propensity score calipers	193
Text matching without propensity score calipers	217

Table 4.3 shows the effective sample size (ESS) of the data before matching, after PSM and after text matching within and without propensity score calipers. The ESS before matching is the minimum of the number of observations in the treatment and

control groups. In our case this is the number of observations in the control group. For PSM and text matching, the effective sample size is the number of matched pairs after matching. In Table 4.3, the ESS for the two different text matching approaches is the average ESS of all text matching methods rounded to the nearest integer. The maximum ESS for text matching with propensity score calipers is 196. The maximum ESS for text matching without propensity score calipers is 223. We observe that the ESS is largest before matching, because no pruning of observations in treatment and control group has happened yet. Out of the matching approaches, matching on text yields larger ESS.

4.3 Causal Effects

The previous section showed that text matching without propensity score calipers performed significantly worse compared to text matching within propensity score calipers. Hence, text matching without propensity score calipers is not considered when looking at causal effect estimation results. First, causal effects with data before matching and data after propensity score matching (PSM) are compared. Table 4.4 above shows the estimated Average Treatment Effect on the Treated (ATT) and standard errors of our outcome variables for both unmatched data and data after PSM. We look at the ATTs and standard errors measured one (t+1) and three years (t+3) after the start year of an arrangement.

Table 4.4: Causal effects of IMF public sector conditions on health outcomes before matching and after PSM

Outcome	ATT	Standard error
Child mortality (t+1) before matching	-7.76	6.08
Child mortality (t+1) after PSM	6.55	7.58
Child mortality (t+3) before matching	-7.52	5.88
Child mortality (t+3) after PSM	6.77	7.26
Health expenditure (t+1) before matching	0.25	0.49
Health expenditure (t+1) after PSM	-0.25	0.52
Health expenditure (t+3) before matching	0.33	0.47
Health expenditure (t+3) after PSM	-0.15	0.47
Vaccination (t+1) before matching	7.25	2.18
Vaccination (t+1) after PSM	-1.20	2.25
Vaccination (t+3) before matching	5.14	2.01
Vaccination (t+3) after PSM	-2.12	2.21

As can be observed in Table 4.4, the effects before matching and after PSM seem to go in different directions for all outcome variables. For example, before matching, the causal effect for the under-five child mortality rate is negative (i.e. child mortality rate is decreasing), whereas after PSM the causal effect is positive (i.e. child mortality rate is increasing). Furthermore, the impacts of child mortality and vaccination rates are stronger when measured three years after treatment, compared to one year. In contrast, the impact of health expenditure is stronger when measured one year after treatment, compared to three years after treatment. However, note that the causal effects on health expenditure are not very reliable, due to the fact that more than 150 values needed to be imputed for this outcome variable. Moreover, the standard error seems relatively large for all measurements. In all measurements after PSM, the standard error is larger in magnitude than the ATT.

Table 4.5: Causal effects of IMF public sector conditions on health outcomes after text matching within propensity score calipers.

Outcome	avg(ATTs)	sd(ATTs)	avg(standard errors)
Child mortality (t+1)	4.85	0.49	6.79
Child mortality (t+3)	5.13	0.50	6.53
Health expenditure (t+1)	-0.23	0.06	0.49
Health expenditure (t+3)	-0.14	0.06	0.45
Vaccination (t+1)	-0.43	0.43	2.20
Vaccination (t+3)	-1.35	0.19	2.07

Table 4.5 shows the average of the ATTs and standard errors, as well as the standard deviation (SD) of the ATTs, of all text matching methods within propensity score calipers. (We will look at the individual ATTs and standard errors of text matching within propensity score calipers subsequently.) Comparing Table 4.5 with Table 4.4, we observe that the causal impacts after text matching goes in the same direction as after PSM, but the magnitude of differences are slightly more conservative. The standard errors are still larger in magnitude than the ATTs, even when accounting for the standard deviation of the ATTs.

The subsequent three figures depict the individual ATTs and standard errors for text matching within propensity score calipers, measured one year after treatment. We only look at outcomes measured one year after treatment, as corresponding graphs measured three years after treatment follow the same patterns. Those graphs are instead presented in Appendix A.

Figure 4.5: Causal effects of IMF public sector conditions on child mortality (t+1) after text matching within propensity score calipers

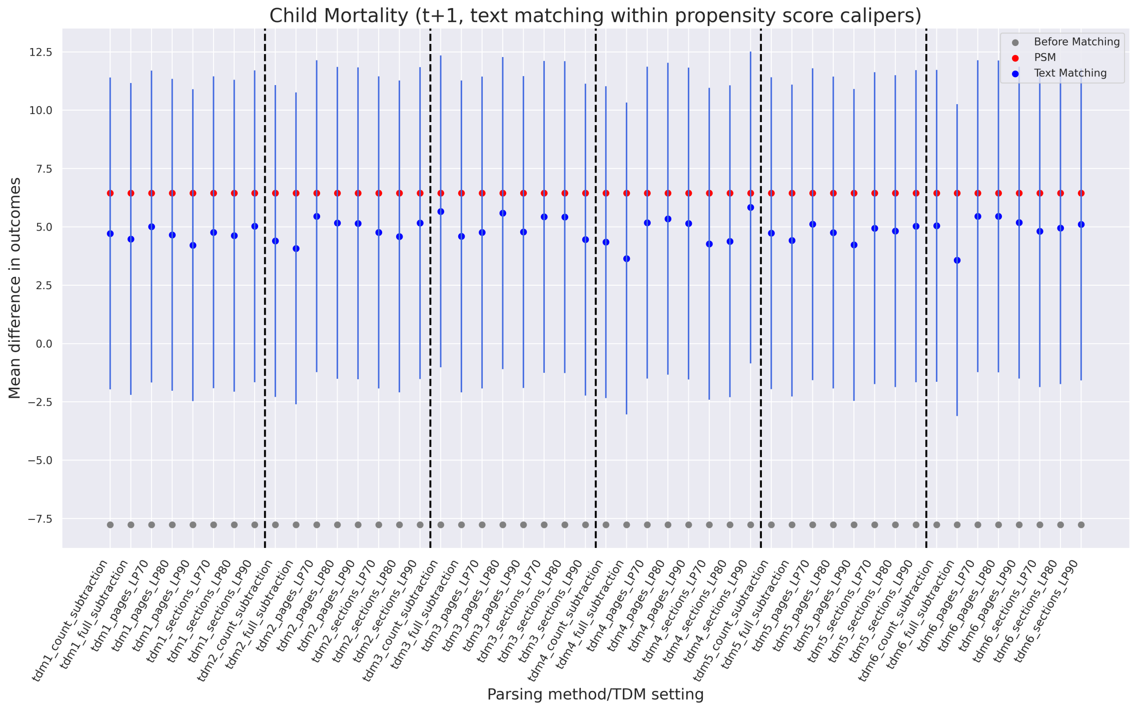


Figure 4.5 above depicts the individual ATTs of child mortality. Gray dots show difference in outcomes before matching, red dots show difference in outcomes after PSM, and blue dots show the ATT for the text matching methods. Standard error bars for the text matching methods are shown as light blue vertical lines. Text matching methods are grouped by TDM setting (see Table 3.5), separated by black dashed lines. As previously observed, the magnitude of child mortality rate for text matching within propensity score calipers is slightly more conservative than PSM with large standard errors. The values have a relatively small spread around 4.85 and no text matching method overlaps with the PSM.

Figure 4.6: Causal effects of IMF public sector conditions on health expenditure (t+1) after text matching within propensity score calipers

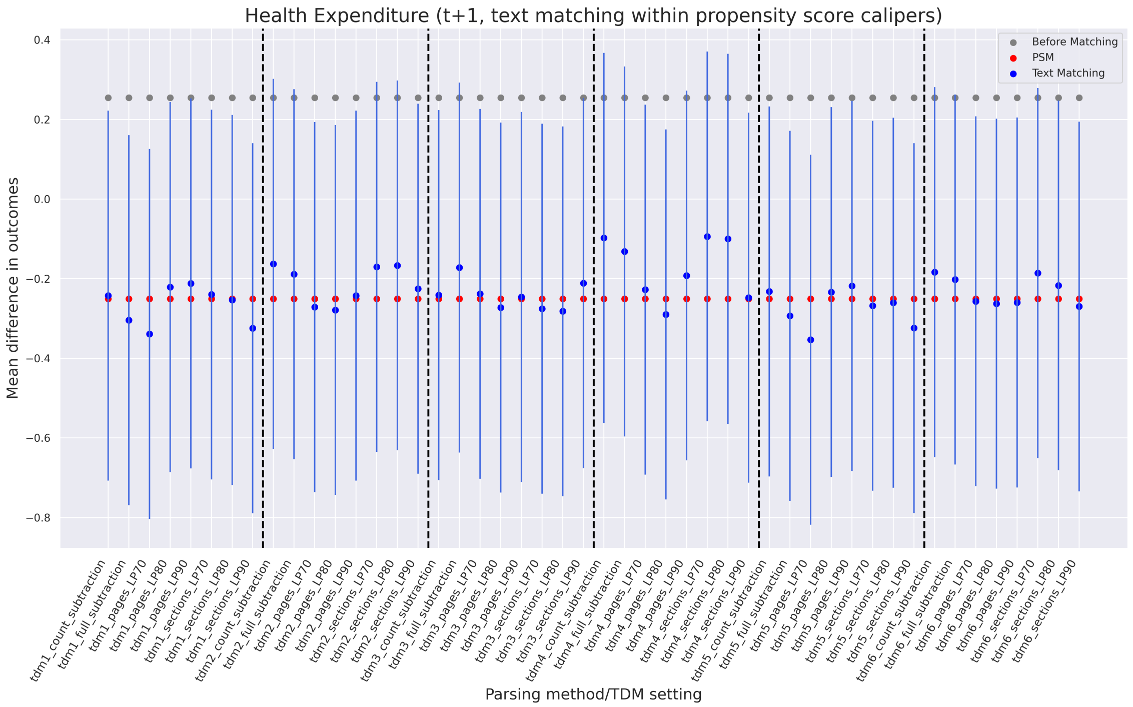


Figure 4.6 above depicts the individual ATTs of health expenditure. Text matching methods overlap with the PSM and the standard errors are large. Some standard errors overlap with the difference in outcomes before matching. Moreover, values of text matching methods with TDM setting tdm4 seem to be closer to the difference in outcomes before matching, compared to text matching methods with other TDM settings.

Figure 4.7: Causal effects of IMF public sector conditions on vaccination (t+1) after text matching within propensity score calipers

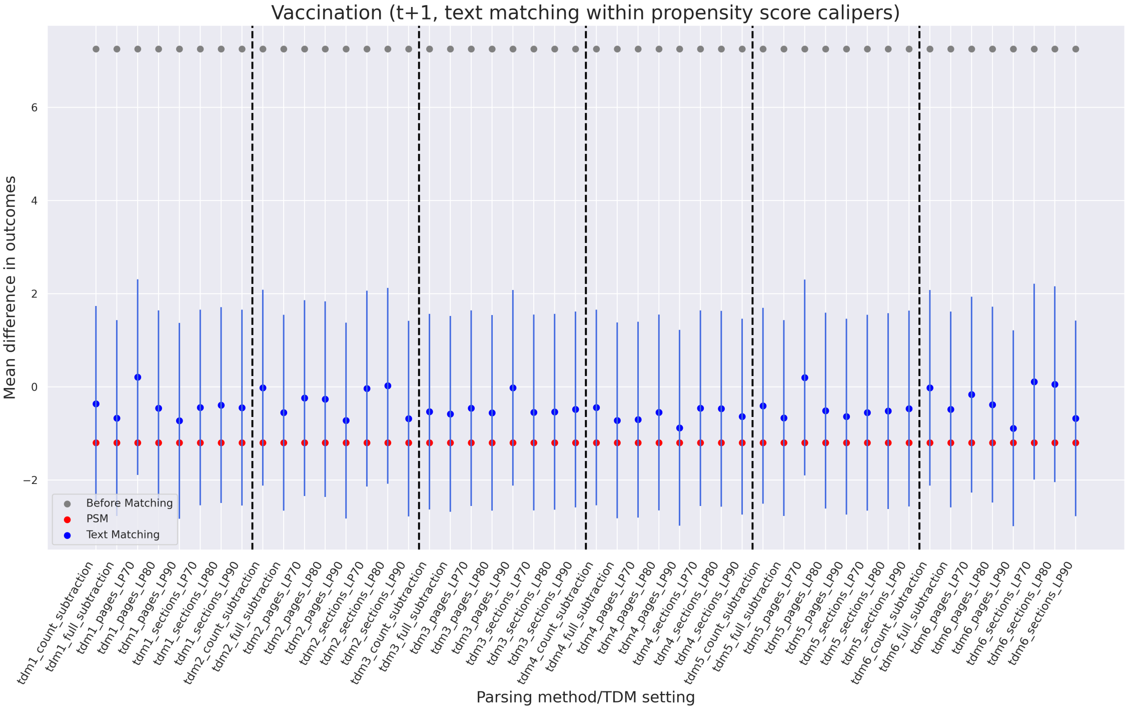


Figure 4.7 above depicts the individual ATTs of vaccination. Similar to child mortality rate, the magnitude of the vaccination rate for text matching within propensity score calipers is slightly more conservative than PSM with large standard errors. Values have a relatively small spread around -0.43 and no text matching method overlaps with the PSM.

5

Discussion

The discussion section first interprets the results. Second, the limitations of the results are discussed. Finally, we make future work recommendations.

5.1 Interpretation of Results

5.1.1 Confounder Parsing

All confounder parsing methods have different trade-offs and the results showed that more than 15% of treatment texts were not mapped to a section in stage one of the confounder parsing procedure. It follows that the amount of treatment text that persists under each text matching method is dependent on the confounder parsing method used. Hence, if one text matching method performs better than another, it might be because the better performing text matching method is able to match on more treatment texts. If this is the case, selection bias persists. On the other hand, if the choice of confounder parsing method is robust under our text matching methods, we see two possible explanations: 1) treatment text does not impact the text matching methods in any significant way and 2) the confounder parsing methods produce equally biased matches after text matching. If the former is true, stage two of our confounder parsing procedure would be superfluous and one would only need to be concerned about how PDFs are converted to raw text. If the latter is true, we are not handling the removal of treatment texts correctly and our confounder parsing procedure would need to be revised.

5.1.2 Covariate Balance

To summarize, we have observed the covariate balance of three different matching approaches:

1. propensity score matching (PSM),
2. text matching without propensity score calipers, and
3. text matching within propensity score calipers.

PSM serves as a baseline that we compare the two text matching approaches against. Key findings show: 1) matching on text covariates can improve covariate balance if used in combination with a propensity score caliper based on the numerical covariates, 2) text matching is relatively robust to tuning TDM parameters and choice of confounder parsing method, and 3) text matching increases the effective sample size.

Disregarding the large standard errors, PSM is able to adequately balance most of the numerical covariates, but fails to sufficiently adjust for differences between treatment and control groups on some of the covariates. Text matching within propensity score calipers shows an overall improvement with regards to covariate balance, and for some of the covariates it is able to considerably reduce the imbalance. This is in line with Mozer et al. (2019), who in their bedside transthoracic echocardiography (TTE) application also found that text matching within propensity score calipers improved the overall covariate balance. Mozer et al. (2019) explains, "when text matching within propensity score calipers, small differences in estimated propensity scores across control units will be offset by any large differences in text." When using a propensity score caliper, we are discarding any treated units for whom the nearest control unit is not within 0.3 standard deviations of the estimated distribution of propensity scores of the numerical confounders we are analysing covariate balance on. Within this space, text-based cosine distance offers a more refined measure of pairwise similarity compared to propensity score distance (Mozer et al., 2019). This allows for more efficient and precise optimization of the resulting matched sample, and is a possible explanation for why we see an improvement in covariate balance, compared to PSM.

In contrast, text matching without propensity score calipers worsened the overall covariate balance, compared to PSM. When text matching without propensity score calipers, these numerical confounders are not considered in the matching procedure. Intuitively, if a numerical covariate cannot be well represented by summary measures of text, it will only be balanced by randomness or by correlation with other covariates that we are analysing, which might explain the poor balance of some covariates. In the context of our data, setting a propensity score caliper when matching on text has therefore shown crucial in order achieve adequate balance on the numerical covariates. However, Mozer et al. (2019) set this caliper to 0.1 standard deviations of the estimated distribution of propensity scores, while we used 0.3 standard deviations, indicating the need for fine-tuning.

Mozer et al. (2019) found TDM-based representations with cosine matching to be relatively robust to tuning parameters, including the choice of bounding and weighting schemes. This agrees with our findings, as the choice of TDM setting does not seem to affect covariate balance in any significant way under our two text matching approaches. Moreover, we also found the choice of confounder parsing method to be robust under our text matching approaches, which is one of our research questions. Because of the different trade-offs each parsing method has, this finding is particularly interesting. As aforementioned, we see two possible explanations for this: treatment text does not impact the text matching methods in any significant way and 2) the confounder parsing methods produce equally biased matches after text matching.

Similar to Mozer et al. (2019), we find that text matching increases the effective sample size (ESS), compared to PSM. This highlights the efficiency of text matching. When text matching within propensity score calipers, we were able to both improve the covariate balance and increase the ESS at the same time. This is important, because the larger our sample is, the less uncertainty we get when evaluating covariate balance and estimating causal effects.

5.1.3 Causal Effects

We observed causal effects of PSM and all text matching methods within propensity score calipers. The large standard errors, which stem from our small sample size, show the uncertainty of our estimates. Disregarding the large standard errors, our findings seem to be in line with the general academic understanding that IMF programs negatively impact the public health of participating countries. Just like Daoud and Reinsberg (2019), our findings also suggest that, on average, IMF programs increase child mortality and decrease health expenditure and vaccination rates in the participating countries. Nooruddin and Simmons (2006) also conclude that IMF programs lead to decreases in health spending and Forster et al. (2019) likewise found that IMF programs increase child mortality rate. Additionally, our causal impacts are less strong than those of Daoud and Reinsberg (2019). This smaller impact may be attributed to the fact that we adjusted for both text and numerical data, whereas Daoud and Reinsberg (2019) only adjusted for numerical data. Moreover, our different results could also stem from Daoud and Reinsberg (2019) having estimated the ATE using instrumental variables, whereas we estimated the ATT using text matching.

For all three outcomes, text matching within propensity score calipers seems to lessen the magnitude of the causal effect, compared to the estimated causal effects that are solely based on propensity score matching. Text matching and PSM being different in magnitude makes intuitive sense, since we have found that adding text matching within propensity score calipers improves covariate balance and thus we adjust for more confounding effects. The more confounding we adjust for, the closer we get to the true treatment impact.

Because text matching within propensity score calipers and PSM yield different results, our findings suggest that IMF policy documents contain valuable information that enables us to control for confounding effects even better, resulting in a more realistic representation of the treatment impact. Text enables us to observe the previously unobservable. In the context of the IMF, we hypothesize one type of confounder that can be observed text and not numerically: political will. Political will can be defined as the willingness of politicians or the government as a whole to make decisions and pursue policies that yield net-benefits for the country in the long-term, regardless of the public opinion on such matters (Przeworski & Vreeland, 2000). A government with strong political will is more likely to improve the health of it's citizens, compared to a government with weak political will. Such a strong government may also be more likely to engage in an IMF program and to successfully implement it. As a result, if we do not control for political will, its impact on public health will be attributed to IMF programs. Other confounders, that are hard to put in numbers, but could be observed in text, are: a government's reputation, negotiation position or the population's trust in the government (Przeworski & Vreeland, 2000).

5.2 Limitations of Results

As explained previously, our small dataset is a consequence of using observations with text. The small dataset reflects itself in our results, as the standard errors are large. Both when evaluating covariate balance and estimating causal effects. The smaller our dataset is, the more uncertainty persists. We observed that the ESS is larger after text matching compared to PSM. However, even after this increase in ESS, our sample size remains small because the original dataset is small. With IMF data, this is a trade-off one needs to make when choosing to match on text. There are only policy documents for those countries that participated in an IMF program. Consequently we do not have data for developed nations (i.e. nations that have not been treated) and are forced to estimate the Average Treatment Effect on the Treated (ATT) instead of the Average Treatement Effect (ATE). On the other hand, if one chooses not to match on text, but only on numerical confounders, then one would have access to significantly more data, because we would not be limited to only country-year observations of treated countries. One would have access to data on country-year observations measured every year from 1980 to 2014 on both developed and developing countries. Hence, an important question arises: is it worth reducing your dataset in order to be able capture confounding that is only observed in text? This is something we leave for future research.

This study was limited to text matching methods with TDM-based text representations in combination with cosine distance metric. However, as aforementioned, other combinations of text representation and distance metrics might yield different results. Hence, the results do not uncover whether covariate balance is robust to confounder parsing in general, but only to this combination of text representation and distance metric. Moreover, we did not exhaust all 14 mapping level thresholds for the section parsing (SP) and page parsing (PP) strategies. For instance, it would be interesting to see how SP with mapping level threshold 14 (1562 cumulative mappings) compares to SP with mapping level threshold 1 (4677 cumulative mappings), with regards to covariate balance. This could be taken even further by investigating how covariate balance is affected by not removing any treatment texts from the policy documents at all, i.e. ignore stage two in the confounder parsing procedure.

Furthermore, the confounder parsing results show how many treatment texts we were able to map to sections, but that does not tell us exactly how many treatment texts we were able to remove from the policy documents. The lower mapping level threshold we use, the less constrained the rule for what a valid mapping is and the more false-positives we get. Furthermore, there is no way of telling whether our confounder parsing methods remove latent treatment text. Our confounder parsing procedure makes the assumption that all treatment texts have been identified and added to the dataset coded by the researchers at the University of Cambridge. Finally, since our contribution is mainly methodological and not substantive, we do not do any significance tests or robustness checks when analyzing causal effects. Therefore, we are unable to make any statements about the causal impacts of the IMF policy conditions with great confidence.

5.3 Future Work Recommendations

We have shown that text matching can effectively improve covariate balance in the field of policy evaluation. Hence, we recommend researchers to use text matching for causal inference in this field and other social science domains. The next step for researchers would be to try out other combinations of text representation and distance metrics to see if the confounder parsing methods are still robust. This could be done by conducting a simulation study where all parameters are controlled for. In such a simulation study one could modify IMF data by manual or automatic methods and purposely insert treatment text of different quantities. One would then analyze how covariate balance is influenced by these different quantities of treatment text.

In text matching, best practice is to iteratively try out different combinations of text representation and distance metric until you find one that produces the best match quality (Mozer et al., 2019). Once this is done, one can expand on our causal effect findings by performing significance tests to ensure the validity of their causal findings, perform sensitivity analysis to assess whether all confounders have been accounted for and robustness tests to check whether one's assumptions are true (Stuart, 2010).

6
Conclusion

We have applied text matching to IMF policy documents and investigated: 1) whether different methods of confounder parsing influence covariate balance differently, and 2) whether text matching would improve covariate balance. In relation to the first research question, we find that the covariate balance is relatively unchanged by the choice of confounder parsing methods. In relation to the second research question, we find that text matching within propensity score calipers based on numerical covariates improves covariate balance, compared to merely using propensity score matching.

In relation to previous research, our findings align with those of Mozer et al. (2019) who found that TDM-based text representations are robust to tuning of parameters and that text matching increases the effective sample size. Moreover, the negative causal effects of IMF labor policy conditions on health, that we find, reflect the findings of the wider academic literature. However, due to large standard errors, our findings have relatively high uncertainty. Just like Daoud and Reinsberg (2019), our findings also suggest that, on average, IMF programs increase child mortality and decrease health expenditure and vaccination rates in the participating countries. Furthermore, we find that text matching within propensity score calipers seems to lessen the magnitude of the causal effect, compared to the estimated causal effects that are solely based on propensity score matching. Hence, our findings suggest that text matching yields more realistic causal effects after having accounted for textual confounders, such as political will, that are not represented in numerical confounders.

One key limitation of our methodology is the small dataset stemming from the fact that we can only use data on countries that have participated in IMF programs. As a result, our standard errors are large. In addition, we did not have the resources to do a human quality analysis of the text matches. We have only tried TDM-based text representations in combination with cosine distance, whereas other combinations may yield different results. Finally, latent treatment text might not have been detected by the confounder parsing methods.

Knowing that the choice of parsing method does not matter, researchers can choose the parsing method that suits their research best, without having to be too concerned with the impacts of the parsing design on covariate balance. They could for example pick the parsing method that allows them to afterwards use the widest range of text representations, giving the most freedom to explore different ways of matching text. Nevertheless, when using different data, the choice of parsing method may have a significant impact on the covariate balance. We therefore urge that researchers consider how similar their use case is to ours when extrapolating from results based on our experiments.

Bibliography

Adobe. (2006). *Adobe® PDF (Portable Document Format) 1.7 Reference*. Retrieved April 26, 2021, from http://archive.org/details/pdf1.7

Atoyan, R., & Conway, P. (2006). Evaluating the impact of IMF programs: A comparison of matching and instrumental-variable estimators. *The Review of International Organizations*, *1*(2), 99–124. https://doi.org/10.1007/s11558-006-6612-2

Barro, R. J., & Lee, J.-W. (2005). IMF programs: Who is chosen and what are the effects?$. *Journal of Monetary Economics*, *52*, 1245–1269.

Bird, G., & Rowlands, D. (2002). Do IMF Programmes Have a Catalytic Effect on Other International Capital Flows? *Oxford Development Studies*, *30*(3), 229–249. https://doi.org/10.1080/1360081022000012671

Brune, N., Garrett, G., & Kogut, B. (2004). THE INTERNATIONAL MONETARY FUND AND THE GLOBAL SPREAD OF PRIVATIZATION, 37.

Butkiewicz, J. L., & Yanikkaya, H. (2005). The Effects of IMF and World Bank Lending on Long-Run Economic Growth: An Empirical Analysis. *World Development*, *33*(3), 371–391. https://doi.org/10.1016/j.worlddev.2004.09.006

Cox, D. R. (1958). *Planning of experiments* [Pages: 308]. Wiley.

Daoud, A., Nosrati, E., Reinsberg, B., Kentikelenis, A. E., Stubbs, T. H., & King, L. P. (2017). Impact of International Monetary Fund programs on child health. *Proceedings of the National Academy of Sciences*, *114*(25), 6492–6497. https://doi.org/10.1073/pnas.1617353114

Daoud, A., & Reinsberg, B. (2019). Structural adjustment, state capacity and child health: Evidence from IMF programmes [Publisher: Oxford Academic]. *International Journal of Epidemiology*, *48*(2), 445–454. https://doi.org/10.1093/ije/dyy251

Daoud, A., Reinsberg, B., Kentikelenis, A. E., Stubbs, T. H., & King, L. P. (2019). The International Monetary Fund's interventions in food and agriculture: An analysis of loans and conditions. *Food Policy*, *83*, 204–218. https://doi.org/10.1016/j.foodpol.2019.01.005

Dreher, A., & Vaubel, R. (2004). The Causes and Consequences of IMF Conditionality. *Emerging Markets Finance and Trade*, *40*(3), 26–54. https://doi.org/10.1080/1540496X.2004.11052571

Edwards, M. S. (2005). Investor Responses to IMF Program Suspensions: Is Noncompliance Costly?*. *Social Science Quarterly*, *86*(4), 857–873. https://doi.org/10.1111/j.0038-4941.2005.00360.x

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). How to Make Causal Inferences Using Texts [arXiv: 1802.02163]. *arXiv:1802.02163 [cs, stat]*. Retrieved December 9, 2020, from http://arxiv.org/abs/1802.02163

Eichengreen, B., Poonam, G., & Mody, A. (Eds.). (2008). *Sudden stops and IMF-supported programs* [OCLC: ocn167489468]. NBER Working Paper Series.

European Union. (2021). European Union. Retrieved May 20, 2021, from https://europa.eu/european-union/index_en

Evrensel, A. Y. (2002). Effectiveness of IMF-supported stabilization programs in developing countries. *Journal of International Money and Finance*, *21*(5), 565–587. https://doi.org/10.1016/S0261-5606(02)00010-4

Explosion. (2021). Spacy: Industrial-strength Natural Language Processing (NLP) in Python. Retrieved May 11, 2021, from https://spacy.io

Forster, T., Kentikelenis, A. E., Stubbs, T. H., & King, L. P. (2019). Globalization and health equity: The impact of structural adjustment programs on developing countries. *Social Science & Medicine*, 112496. https://doi.org/10.1016/j.socscimed.2019.112496

Frangakis, C. E., & Rubin, D. B. (2002). Principal Stratification in Causal Inference [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.2002.00021.x]. *Biometrics*, *58*(1), 21–29. https://doi.org/https://doi.org/10.1111/j.0006-341X.2002.00021.x

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental Versus Experimental Estimates of Earnings Impacts [Publisher: SAGE Publications Inc]. *The ANNALS of the American Academy of Political and Social Science*, *589*(1), 63–93. https://doi.org/10.1177/0002716203254879

Goldstein, M. (2002). IMF Structural Conditionality: How Much Is Too Much? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.300885

Greenland, S. (2003). Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias [Publisher: Lippincott Williams & Wilkins]. *Epidemiology*, *14*(3), 300–306. Retrieved May 15, 2021, from https://www.jstor.org/stable/3703850

Greifer, N. (2021). Cobalt: Covariate Balance Tables and Plots. Retrieved April 23, 2021, from https://CRAN.R-project.org/package=cobalt

Guglielmetti, Y. S. +. P. (2020). Pdfminer.six: PDF parser and analyzer. Retrieved April 16, 2021, from https://github.com/pdfminer/pdfminer.six

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT [Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214504000000647]. *Journal of the American Statistical Association*, *99*(467), 609–618. https://doi.org/10.1198/016214504000000647

Hansen, B. B., Fredrickson, M., Buckner, J., Errickson, J., Solenberger, A. R. a. P., & Tseng, w. e. F. c. d. t. D. P. B. a. P. (2019). Optmatch: Functions for Optimal Matching. Retrieved April 23, 2021, from https://CRAN.R-project.org/package=optmatch

Hansen, B. B., & Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows [Publisher: Taylor & Francis]. *Journal of Computational and Graphical Statistics*, *15*(3), 609–627. https://doi.org/10.1198/106186006X137047

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching As An Econometric Evaluation Estimator. *The Review of Economic Studies*, *65*(2), 261–294. https://doi.org/10.1111/1467-937X.00044

Hill, J. L., Reiter, J. P., & Zanutto, E. L. (2004). A Comparison of Experimental and Observational Data Analyses [Section: 5 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1 *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives* (pp. 49–60). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470090456.ch5

Hills, C. A., Peterson, P. G., & Goldstain, M. (1999). *Safeguarding Prosperity in a Global Financial System: The Future International Financial Architecture*. Council on Foreign Relations; Institute for International Economics.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Honaker, J., King, G., & Blackwell, M. (2019a). Amelia: A Program for Missing Data. Retrieved April 23, 2021, from https://CRAN.R-project.org/package=Amelia

Honaker, J., King, G., & Blackwell, M. (2019b). Mi.meld function - RDocumentation. Retrieved April 24, 2021, from https://www.rdocumentation.org/packages/Amelia/versions/1.7.6/topics/mi.meld

Imbens, G. W. (2004). Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, *86*(1), 4–29. https://doi.org/10.1162/003465304323023651

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

International Monetary Fund. (2020). *IMF Annual Report* (tech. rep.).

Ivanova, A., Mourmouras, A., AMourmouras@imf.org, AIvanova@imf.org, Anayotos, G. C., GAnayotos@imf.org, Mayer, W., & WMayer@imf.org. (2003). What Determines the Implementation of IMF-Supported Programs? *IMF Working Papers*, *3*(8), 1–47. https://doi.org/10.5089/9781451842531.001

Jensen, N. M. (2004). Crisis, Conditions, and Capital: The Effect of International Monetary Fund Agreements on Foreign Direct Investment Inflows. *Journal of Conflict Resolution*, *48*(2), 194–210. https://doi.org/10.1177/0022002703262860

Keith, K. A., Jensen, D., & O'Connor, B. (2020). Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates [arXiv: 2005.00649]. *arXiv:2005.00649 [cs]*. Retrieved November 12, 2020, from http://arxiv.org/abs/2005.00649

King, G., Nielsen, R., Coberley, C., Pope, J. E., & Wells, A. (2011). Comparative Effectiveness of Matching Methods for Causal Inference.

Markoulidakis, A., Taiyari, K., Holmans, P., Pallmann, P., Busse-Morris, M., & Griffin, B. A. (2020). *A tutorial comparing different covariate balancing methods* *with an application evaluating the causal effect of exercise on the progression of Huntington's Disease.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality [arXiv: 1310.4546]. *arXiv:1310.4546 [cs, stat]*. Retrieved January 29, 2021, from http://arxiv.org/abs/1310.4546

Mozer, R., Miratrix, L., Kaufman, A. R., & Anastasopoulos, L. J. (2019). Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality [arXiv: 1801.00644]. *arXiv:1801.00644 [cs, stat]*. Retrieved November 10, 2020, from http://arxiv.org/abs/1801.00644

Nooruddin, I., & Simmons, J. W. (2006). The Politics of Hard Choices: IMF Programs and Government Spending. *International Organization*, *60*(4), 1001–1033. https://doi.org/10.1017/S0020818306060334

Palmer, J. A. (2021). Pdftotext: Simple PDF text extraction. Retrieved April 22, 2021, from https://github.com/jalan/pdftotext

Polshcha, M. (2021). Pdfreader: Pythonic API for parsing PDF files. Retrieved April 22, 2021, from http://github.com/maxpmaxp/pdfreader

Przeworski, A., & Vreeland, J. R. (2000). The effect of IMF programs on economic growth. *Journal of Development Economics*, *62*, 385–421.

Rosenbaum, P. R. (1991). A Characterization of Optimal Designs for Observational Studies [_eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1991.tb01848.x]. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 597–610. https://doi.org/https://doi.org/10.1111/j.2517-6161.1991.tb01848.x

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, *79*(387), 516–524. https://doi.org/10.1080/01621459.1984.10478078

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.1080/00031305.1985.10479383

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1975). Bayesian inference for causality: The importance of randomization. *The Proceedings of the social statistics section of the American Statistical Association* (pp. 233–239).

Rubin, D. B. (1976). Inference and missing data. *63*(3), 581–592.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate [Publisher: American Educational Research Association]. *Journal of Educational Statistics*, *2*(1), 1–26. https://doi.org/10.3102/10769986002001001

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *6*(1), 34–58. Retrieved February 3, 2021, from https://www.jstor.org/stable/2958688

Rubin, D. B. (1979). Discussion of "Conditional independence in statistical theory" by AP Dawid. *41*, 27–28.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment [Publisher: [American Statistical Association, Taylor & Francis, Ltd.]]. *Journal of the American Statistical Association*, *75*(371), 591–593. https://doi.org/10.2307/2287653

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B., & Thomas, N. (1996). Matching Using Estimated Propensity Scores: Relating Theory to Practice [Publisher: [Wiley, International Biometric Society]]. *Biometrics*, *52*(1), 249–264. https://doi.org/10.2307/2533160

Schwartz, S., Gatto, N. M., & Campbell, U. B. (2012). Extending the sufficient component cause model to describe the Stable Unit Treatment Value Assumption (SUTVA). *Epidemiologic Perspectives & Innovations*, *9*(1), 3. https://doi.org/10.1186/1742-5573-9-3

Shinyama, Y., Guglielmetti, P., & Marsman, P. (2019). Converting a PDF file to text — pdfminer.six 20201018 documentation. Retrieved April 20, 2021, from https://pdfminersix.readthedocs.io/en/latest/topic/converting_pdf_to_text.html

Simmons, B. A. (2000). International Law and State Behavior: Commitment and Compliance in International Monetary Affairs. *American Political Science Review*, *94*(4), 819–835. https://doi.org/10.2307/2586210

Smith, R. (2021). Tesseract-ocr/tessdoc [original-date: 2020-01-30T08:36:39Z]. Retrieved April 16, 2021, from https://github.com/tesseract-ocr/tessdoc

Steinwand, M. C., & Stone, R. W. (2008). The International Monetary Fund: A review of the recent evidence. *The Review of International Organizations*, *3*(2), 123–149. https://doi.org/10.1007/s11558-007-9026-x

Stone, R. W. (2004). The Political Economy of IMF Lending in Africa. *American Political Science Review*, *98*(4), 577–591. https://doi.org/10.1017/S000305540404136X

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science : a review journal of the Institute of Mathematical Statistics*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes [Place: US Publisher: American Psychological Association]. *Developmental Psychology*, *44*(2), 395–406. https://doi.org/10.1037/0012-1649.44.2.395

Stubbs, T., Kentikelenis, A., Stuckler, D., McKee, M., & King, L. (2017). The impact of IMF conditionality on government health expenditure: A cross-national analysis of 16 West African nations. *Social Science & Medicine*, *174*, 220–227. https://doi.org/10.1016/j.socscimed.2016.12.016

Stuckler, D., King, L. P., & Basu, S. (2008). International Monetary Fund Programs and Tuberculosis Outcomes in Post-Communist Countries (M. Murray, Ed.). *PLoS Medicine*, *5*(7), e143. https://doi.org/10.1371/journal.pmed.0050143

Teorell, J., Dahlberg, S., Holmberg, S., Rothstein, B., Alvarado Pachon, N., & Axelsson, S. (2020). QoG Standard Dataset 2020 [type: dataset]. https://doi.org/10.18157/QOGSTDJAN20

United Nations. (2021). United Nations [Publisher: United Nations]. Retrieved May 20, 2021, from https://www.un.org/en/?

A

Appendix 1

Figure A.1: Causal effects of IMF public sector conditions on child mortality (t+3) after text matching within propensity score calipers

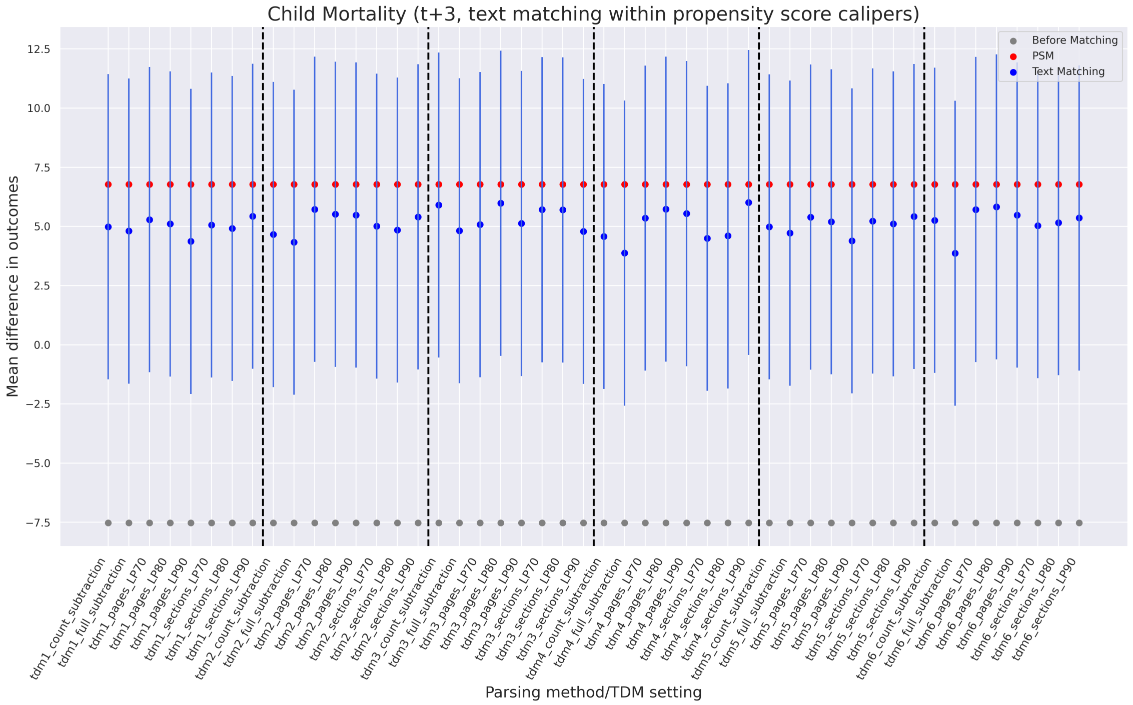


Figure A.2: Causal effects of IMF public sector conditions on health expenditure (t+3) after text matching within propensity score calipers

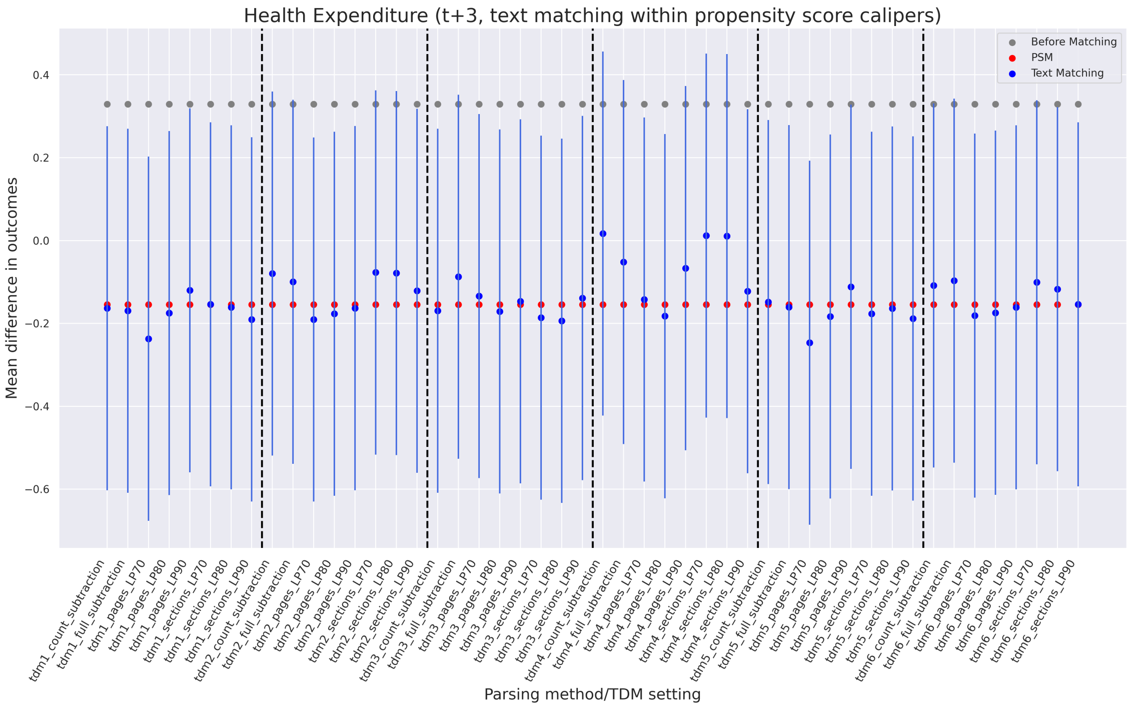
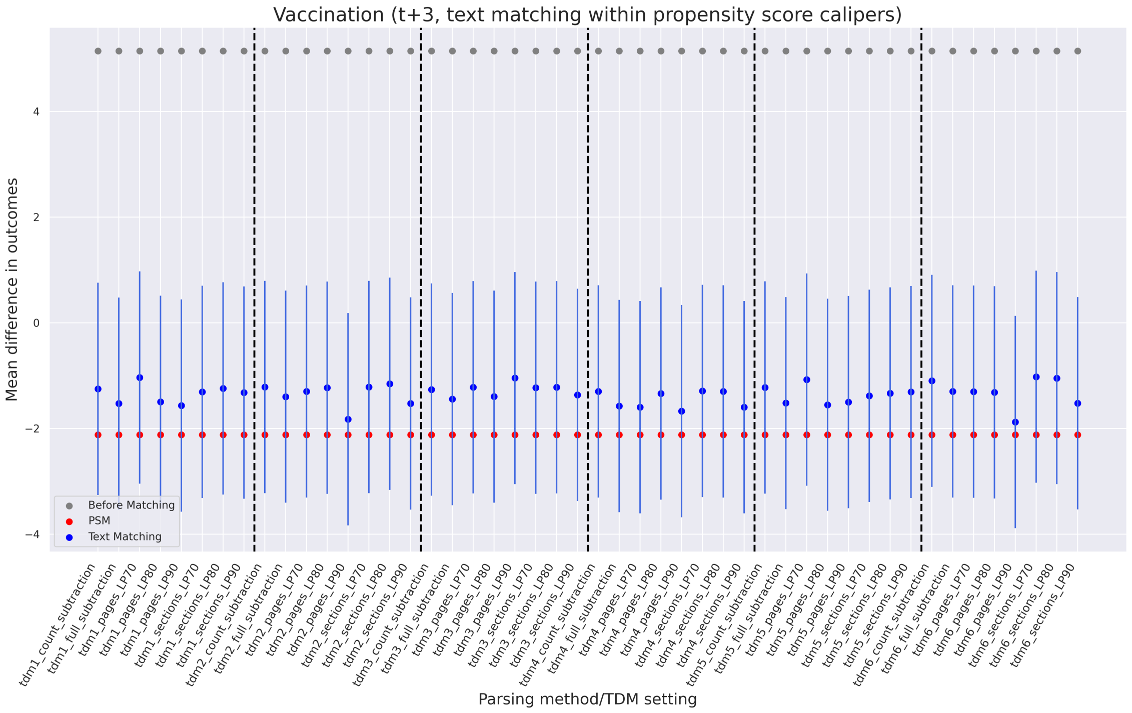


Figure A.3: Causal effects of IMF public sector conditions on vaccination (t+3) after text matching within propensity score calipers


Vaccination (t+3, text matching within propensity score calipers)
Before Matching
PSM
Text Matching
Mean difference in outcomes
Parsing method/TDM setting
tdm1_count_subtraction
tdm1_full_subtraction
tdm1_pages_LP70
tdm1_pages_LP80
tdm1_pages_LP90
tdm1_sections_LP70
tdm1_sections_LP80
tdm1_sections_LP90
tdm2_count_subtraction
tdm2_full_subtraction
tdm2_pages_LP70
tdm2_pages_LP80
tdm2_pages_LP90
tdm2_sections_LP70
tdm2_sections_LP80
tdm2_sections_LP90
tdm3_count_subtraction
tdm3_full_subtraction
tdm3_pages_LP70
tdm3_pages_LP80
tdm3_pages_LP90
tdm3_sections_LP70
tdm3_sections_LP80
tdm3_sections_LP90
tdm4_count_subtraction
tdm4_full_subtraction
tdm4_pages_LP70
tdm4_pages_LP80
tdm4_pages_LP90
tdm4_sections_LP70
tdm4_sections_LP80
tdm4_sections_LP90
tdm5_count_subtraction
tdm5_full_subtraction
tdm5_pages_LP70
tdm5_pages_LP80
tdm5_pages_LP90
tdm5_sections_LP70
tdm5_sections_LP80
tdm5_sections_LP90
tdm6_count_subtraction
tdm6_full_subtraction
tdm6_pages_LP70
tdm6_pages_LP80
tdm6_pages_LP90
tdm6_sections_LP70
tdm6_sections_LP80
tdm6_sections_LP90