1 Introduction

1.1 Motivation and Aims of the Dissertation

One of the major challenges of empirical research is to ensure unbiased results. A common reason for biased results is correlations between independent variables and unobserved factors that also influence the dependent variable. These unobserved factors are referred to as omitted variables, and the resulting bias is known as omitted-variable bias. A common approach in empirical studies to mitigate an omitted-variable bias is to include fixed effects in panel data models. This dissertation demonstrates that a more comprehensive clustered[1] modeling is often necessary to obtain unbiased results. Clustered modeling refers to the combination of a clustering approach and an econometric model (Bijak and Thomas (2012), Bakoben et al. (2020)). The econometric model is optimized or trained with respect to the different data clusters.[2]

One reason for the occurrence of omitted variables is that they may be unavailable or unobservable. To control for such unobservable effects is a fundamental challenge in empirical research. One possible approach is to utilize observable data clusters that reflect unobserved characteristics. For instance, consider the task of predicting the income development of employees based on individual traits, such as personality. Empirical results might suggest that income grows faster for more outgoing and positive-minded employees than for more reserved workers. However, the mood and behavior of the employee may be influenced by firm-specific characteristics, such as management quality, workplace atmosphere or organizational peculiarities. Employees in "high-quality" firms may therefore appear more positive-minded simply because of better working conditions. In such a case, it is not the employees' mood but rather the firm's quality that determines income development. When granular firm-level factors cannot be explicitly modeled, it is necessary to control for aggregated firm-level effects. If these firm clusters are not considered, the results are biased.

[1] Clustering refers to the process of grouping a set of objects in a way that the objects in the same group are more similar than objects in different groups (e.g., Bosker et al. (2024)).

[2] A common approach is to first cluster the data and then to optimize the models on each cluster separately. For more information on clustered modeling, see Chapter 2.

Consequently, the consideration of appropriate data clusters is vital to obtain unbiased models. However, as with omitted variables, these clusters may themselves be unavailable or unobservable. For example, it may not suffice to consider firm-specific characteristics. But it may be necessary to account for department-specific characteristics. If no information at department level is available, the resulting estimates are likely to be biased.

We[3] find in the literature across various fields of research that unconsidered data clusters are a common reason for biased results. In many cases, this can be attributed to a lack of sufficient information to perform the required clustering. Against this background, the aim of the dissertation is two-fold: First, we illustrate that such unconsidered clusters currently lead to misleading and biased results in the literature. Second, we introduce alternative clustered models that enable the identification and consideration of these clusters. As a result, we show that such clustering approaches are essential in empirical economics and must be considered in empirical models.

1.2 | Course of Investigation

The dissertation is divided into four main chapters. The first main chapter gives a theoretical overview of the topic of clustered modeling with respect to the approaches employed in this dissertation. Each of the remaining three main chapters covers a different area of research. It follows a brief description of each research question, the shortcomings regarding the choice of clusters and our alternative approach.

Chapter 3 covers the field of Labor Economics and investigates how employees are affected by cultural differences when migrating internationally. A common approach is to cluster employees in "migrants" and "natives". With this method, one typically observes that the "migrants" are at a disadvantage and must first adapt to the new labor market.[4]

But with this approach it is not possible to consider various employee clusters. For instance, any job change, regardless of cultural background, requires adaptation. International migration typically coincides with a change in employer, position, or both, making it necessary to isolate culture effects from general job change effects. This results in the necessity to consider clusters of workers that change jobs or positions. Besides these clusters, clusters of returning, internationally experienced or high-skilled employees must be considered. This requires a level of employee information that the employed data in the literature cannot provide.

[3] The following chapters are based on joint publications. I will therefore refer to "we".

[4] This is a very compact summary, and there are many exceptions. Please refer to Sections 3.1 and 3.2 for more detailed explanations.

As an alternative, we utilize employment data from professional soccer. With this approach, we are able to include all employee clusters mentioned. These clusters are considered in the form of mixed effects models and interaction effects with dummy variables.[5]

We can utilize a closed, well documented system of employee movements and an extensive data base of employee performance and valuation. We observe that overall, i.e., without consideration of the mentioned clusters, cultural differences due to a club change negatively affect employee value in the short term. But when we consider the clusters and isolate the culture effect from general job change effects, we can observe positive culture effects in the short and the long term.

In Chapter 4, we face a similar situation with relevant data clusters that are not considered in the empirical model. It covers the famous uncertainty of outcome hypothesis. The hypothesis states that interest in a sports game depends on how uncertain the potential audience is about the outcome. According to the hypothesis, games in which both teams have similar chances of winning attract a larger audience and generate more interest, as the games are expected to be more exciting. The literature finds evidence in favor of the hypothesis and against the hypothesis. The empirical analysis is typically based on the number of stadium attendees or the size of the TV audience. But in both cases, the audience is largely anonymous. Consequently, it is not possible to cluster the audience into appropriate groups. For example, it is obvious that the home team fans and away team fans react differently to a high winning probability of the home team.

We take an alternative approach and analyze the number of comments a game receives on social media. With this method, we can cluster the audience based on their commenting behavior. We optimize a regression model on each of these data clusters. We find that home fans, away fans and neutrals each favor specific game outcomes and that they try to avoid undesirable outcomes. Therefore, they engage in a form of loss aversion. Only if we account for this form of loss aversion in the different fan clusters, can we isolate and confirm the uncertainty of outcome hypothesis. Otherwise, the overall effect depends on the sizes of the fan clusters and their behavior in response to loss aversion.

Chapters 3 and 4 cover two similar cases that highlight the necessity of appropriate clustering. But the question remains whether some form of "optimal" or "best-in-class" clustered model exists. To answer this question, Chapter 5 deals with the topic of credit risk modeling. The aim is to achieve a model with a high prediction accuracy of the occurring losses in case of a default. It is necessary to segment the borrowers

[5]Chapter 2 categorizes different forms of clustered modeling and explains that mixed effects models can be considered a form of clustered modeling.

to predict accurate losses. But the segmentation can be based on multiple borrowers and contract properties. This includes the asset class, number and type of collaterals, number of guarantors, industry or country. We introduce a method to segment the data in such a way that the employed machine learning methods perform best on these clusters compared to other clustering methods. The approach is to select the most important variables with Shapley values. These variables are employed to cluster the data. On each data segment, a machine learning model is trained. We find that the type of clustering approach is vital to achieve the highest predictive power possible.

Summarized, we observe from all three cases that appropriate clustered modeling is essential to ensure unbiased results and a high predictive accuracy. We find that these are currently not adequately considered in the literature. The dissertation will be concluded by Chapter 6 with a summary of the most important results.

2 The Use of Clusters for Econometric Modeling

2.1 Theoretical Background

In the following chapters, it will be necessary to cluster or segment the data to achieve unbiased and accurate models. Clustering refers to the process of grouping a set of objects in a way that the objects in the same group are more similar than objects in different groups (c.f. for example Bosker et al. (2024)). Clustered modeling[1] refers to the combination of a data clustering approach and an econometric model (Bijak and Thomas (2012), Bakoben et al. (2020)). The econometric model is optimized or trained with respect to the different data clusters. In this chapter, we give a theoretical overview of clustered modeling and categorize the approaches employed in the following chapters.

Data can be clustered based on different segmentation bases and segmentation methods (Wedel and Kamakura (2000), Bijak and Thomas (2012)). The segmentation basis refers to a set of variables that allow us to assign observations to homogeneous groups (Bijak and Thomas (2012)). These variables can be either observable or unobservable (Wedel and Kamakura (2000)). Unobservable variables may lead to ineffective segmentation[2] since they cannot be directly used by the segmentation method. This leads to overlapping clusters with little discriminative power. To get a segmentation basis that allows for effective segmentation proves to be one of the main challenges in Chapter 3 and Chapter 4.

Segmentation methods can either be based on statistical clustering approaches or on experience-based heuristics (Siddiqi (2005), Bijak and Thomas (2012)). In many cases, the necessary clusters are intuitive. For example, in a data set with only a few different companies it may be necessary to segment the data for each company. This is a case of experience-based clustering. But if there are too many companies in relation

[1] Alternative names for clustered modeling include segmented modeling, clustering and segmentation.

[2] Wedel and Kamakura (2000) define six criteria for effective segmentation. The most important ones are identifiability and stability: It must be possible to assign observation to segments and the obervations in the segments must differ from each other.

to the number of observations, it may be necessary to employ a statistical approach to aggregate several companies into a single cluster to prevent overfitting.

Once the segmentation basis and method are defined, the clustered model can use either a two-step or a simultaneous approach (Aurifeille (2000), Bijak and Thomas (2012)). In a two-step approach, the segmentation is performed first and the econometric model is employed on each segment. In a simultaneous approach, the clusters and the econometric model are optimized at the same time. For the example of a linear regression model, both modeling approaches lead to regression coefficients that differ between the clusters.[3] For a vector of clusters $C = c_1, \ldots, c_m$, a dependent variable $Y_{i(c_j)}$ $(i = 1, \ldots, n)$, independent variables $(\mathbf{X})_{i(c_j)}$, regression coefficients β_{c_j} and the residuals $\varepsilon_{i(c_j)}$, the clustered model can be formulated in the following way. The notation $i(c_j)$ indicates that observation i belongs to cluster c_j:

$$Y_{i(c_j)} = \beta_{0,c_j} + \beta_{c_j}^{'} X_{i(c_j)} + \varepsilon_{i(c_j)}, \qquad j = 1, \ldots, m \tag{2.1}$$

This implies that for each cluster the intercept and the slope of the regression coefficients can differ. In this example, the regression model can be optimized by optimizing the model on each cluster c_j separately. An alternative approach is to include dummy variables that encode the cluster associations:

$$I_j = (0, \ldots, 0, \underbrace{1}_{\text{j-th position}}, 0, \ldots, 0) \tag{2.2}$$

$$Y_i = I_{j(i)}\beta_{0,C} + I_{j(i)}\beta_C^{'} X_i + \varepsilon_i \tag{2.3}$$

Similar to the definition above, $j(i)$ indicates that the cluster association is derived from i. The matrix β_C includes all regression coefficients $\beta_{c_1}, \ldots, \beta_{c_m}$. Therefore, in this example the multiple model approach (one regression model for each cluster) can be transformed into a single model where clusters are implicitly modeled with dummies. One advantage of the multiple model approach is that the results are usually more intuitive. Additionally, it can employ different models, variable sets or hyperparameters on each segment (see Chapter 5).

2.2 Overview of the Employed Clustered Modeling Approaches

In the following, we explain and compare the employed clustered modeling approaches in the subsequent chapters based on the definitions in Section 2.1. All approaches utilize

[3] We assume that each observation is in exactly one cluster.

a two-step approach: First, the required clusters are defined. Second, the model is optimized or trained with respect to these clusters.

Chapter 3 analyzes the effect of cultural differences on employees that change their job internationally. We find[4] that this requires the consideration of various employee segments: This includes a segmentation based on the job (change) situation and clusters of returning or internationally experienced employees. Additionally, the model must control for high-performing employees that are more likely to change internationally. To identify these employee clusters, granular information about the employment and performance history of employees is necessary that is usually unavailable with "traditional" approaches. Instead, to achieve an effective segmentation, soccer data is employed in Chapter 3 that provides the required information.

Since the segmentation is intuitive, the clusters are defined experience-based. The empirical model utilizes a mixed effects model and dummy variables to consider the data segments in the model. The mixed effects model enables the regression coefficients to deviate between the defined clusters and is for example used by Saeed et al. (2018) for this purpose. We compare the results with an unclustered model, i.e., with a model that does not consider the employee segments mentioned. We find only in the unclustered model significant negative effects of cultural differences on the employee value[5] over time. One reason is that the culture effect is not isolated from job change effects in the unclustered model. Contrarily, the clustered model is able to isolate job change effects from culture effects and shows generally only significant positive long-term and under certain circumstance short-term effects.

Chapter 4 concerns the uncertainty of outcome hypothesis and loss aversion. The uncertainty of outcome hypothesis states that interest in a sports game depends on the uncertainty of the outcome of a game. Loss aversion refers to the tendency of fans to avoid interacting with games with a high loss probability of their favorite team. We find that it is essential to segment the audience in home and away fans and neutrals to analyze the effects. Commonly, the literature employs for the interest either the number of stadium attendees or the size of the TV audience. With both measures, the audience is anonymous and doesn't allow for the required segmentation. Instead, the number of comments on social media are employed in Chapter 4 to determine the interest in a game. This approach allows for an effective segmentation that distinguishes between the fan groups. The segmentation is performed based on the number of comments of a commenter on games of different teams. The clustering approach is again experience-

[4]For more information about all studies including the motivation and results, refer to the relevant Chapters 3-5.

[5]Value refers to the market value of the soccer player.

based with some statistical justification to distinguish neutrals from fans. In contrast to Chapter 3, separate fixed effects models are employed on each of the fan clusters. We find different effects of result uncertainty on the clusters. Consequently, the total effect across all fan groups depends on the fan distribution in the audience.

In Chapter 3 and 4 the main challenge is the segmentation basis. For the clustered model, largely experience-based clustering methods and linear regression methods were employed. In comparison, Chapter 5 utilizes several different statistical segmentation methods and various machine learning methods and aims to identify the best approach for the case of LGD modeling. For this purpose, an unclustered model is compared with several clustered modeling approaches. The clustered modeling approaches utilize different segmentation methods. In each case, different machine learning methods are employed on each of the identified clusters. We find that the best clustered modeling approach significantly outperforms the unclustered model.

3 The Impact of Cultural Differences on the Success of Elite Labor Migration – Evidence from Professional Soccer[1]

3.1 Introduction

We find over the course of this dissertation that clustered modeling approaches are necessary to isolate effects of interest in various fields of research. As already explained in Section 2.2, one of the main challenges is often the availability of the required data for the segmentation. With "traditional" approaches, relevant segments can often not be identified. Because of its inherent complexity and extensive data requirements, the topic of international migration is a prime example. We begin by illuminating the research background and by describing the existing results. Thereafter, we compare an unclustered[2] model (that corresponds with the literature results) with our clustered modeling approach. We find that the unclustered model is biased. The clustered model does not show these distortions.

International migration is an ongoing and prevalent theme in labor economics (Hatton (2014)) and human resource management (Latukha et al. (2022), Healy and Oikelome (2007)) research. Besides the effects of migration on the countries of origin and destination, the effects of international migration on the moving workers' valuation are particularly interesting for the application and hiring decisions of workers and employers.[3]

The literature predominantly identifies negative effects of international migration on moving employees' valuation (Hatton (2014)). These effects are attributed to cultural and linguistic differences (Peeters et al. (2021), Åslund et al. (2014), Kónya (2007)), the need to adapt education and labor experience to the destination country (Weiss et al.

[1] This chapter is based on the study "The Impact of Cultural Differences on the Success of Elite Labor Migration – Evidence from Professional Soccer", published in the Annals of Operations Research in 2024 (Bosker and Gürtler (2024)).

[2] The results of the unclustered model are called total culture effects in the published study.

[3] With employee value we refer to the monetary value the labor market is willing to pay for employees. The cited literature mostly utilizes employees' wages. For more details on how we measure the value in the present study, see Section 3.5.

(2003), Friedberg (2000)), and ethnic discrimination (as observed in Carruthers and Wanamaker (2017), Heywood and Parent (2012)). The literature focusing on asymmetric effects of international migration is an exception to these negative effects: employees moving from "poor" to "rich" countries tend to increase their value because they can better realize their potential in a stronger economy (Yashiv (2021), Becker and Ferrara (2019), Hatton (2014)).

Therefore, a prevalent explanation in the literature (Peeters et al. (2021), Åslund et al. (2014), Kónya (2007)) and the public perception (Alesina et al. (2022)) attributes the negative effects to workers difficulties to adapt to the culture of the destination country. But it turns out that the literature does not fully isolate the culture effect and lacks consideration of various data clusters and individual- and firm-level effects. First, international migration often involves a change of employer and job, which affects the employee's valuation because he or she must adapt to the new employer or job. Therefore, the culture effect analyzed in the literature is not isolated from other adaptation effects and is biased by effects that can be observed due to any employer change, even if it occurs within the same country or firm. Such clusters of workers that change employers or jobs must be considered in the analysis. Second, the effects likely vary depending on employee clusters that differ in their previous migration experiences. For instance, culture effects differ between employees who frequently migrate or return to a country they already worked in and employees who never migrated before. Third, the literature tends to analyze average effects over time. However, cultural adaptations occur gradually with varying effects depending on the time elapsed since migration. Fourth, considering employee clusters that differ in prior job performance is indispensable because employees with extraordinary job performance records tend to change employers internationally. Contrarily, employees with stable jobs and unexceptional performance records tend to remain in the (domestic) cultural group. Therefore, by neglecting past job performance in the analysis, we would measure a biased culture effect because of mixed culture change and job performance effects.[4] In this context, the literature considers as employee characteristics at best work experience at the time of migration, educational level and survey-based subjective performance assessments. Fifth, we must consider the asymmetric effects of international migration. To this end, we control with a metric for the difference in the "strength" of the respective labor markets to capture the effects of (in strength) asymmetric job changes.

Our aim is to model the culture effect in such a way that we can consider the

[4]We especially account for potential effects of selective migration (such as with H-1B visas, see Peri et al. (2015)). In this context, cultural differences between employee and employer, and performance are highly correlated. Without extensive performance measures, the value development can not be attributed to culture effects.