Predicting salaries in the Greater Toronto Area: A Bayesian approach

by

Andres Danilo Castillo Vega

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science
Department of Civil and Mineral Engineering
University of Toronto

Abstract

Labour markets and transportation systems are at the core of urban life. Attributes such as residential and work locations, household income, and auto ownership are closely related to the interactions in the labour market. Since salaries are a key component of these interactions, predicting salaries becomes an important task for integrating labour market outcomes into travel behaviour modelling. Using a data-driven approach, we estimate a Bayesian hierarchical model to predict salary distributions in the Greater Toronto Area. The results of this work demonstrate that this approach provides better estimates at both the aggregated and disaggregated levels and generates more robust predictions by producing probability distributions instead of point estimates. This characteristic is key for using this model in an urban simulation setting such as the ILUTE framework.

Acknowledgements

The completion of this thesis marks the end of a long journey that began at least seven years ago as a dream. This journey has been full of learnings and experiences that define who I am. Today also marks the beginning of a new stage in my life. For that reason, I would like to thank all the people who have been part of this journey and made it possible.

To my love, Ale, for being my support and inspiration. I feel so fortunate to have you in my life. This achievement is also yours. To my love, Emi, you arrived in our lives at the perfect moment, and we are happily waiting for you. I hope someday you can find this work as an inspiration to follow your dreams and never give up. There are no limits if you work with love and dedication.

To Prof. Eric Miller, for his patience, guidance, and support during this journey. I am very grateful for the opportunity to work with you and learn from your experience.

To Prof. Matthew Roorda, for kindly agreeing to be the second reader of this thesis and providing his feedback. Also, to all my colleagues at the University of Toronto, I learned a lot from all of you. Thanks for the good times and memories.

To my family, for their encouragement during this journey. All of you are continuously teaching me about the important things in life. This thesis is dedicated to the memory of my mother, Lucinda, who always believed in me and inspired me to follow my dreams. I am sure you would be very proud of this achievement. All that I have done is thanks to you.

Last but not least, to my dog, Avena. This thesis is also yours because you stayed with me during the long hours of work and study. You do not know how much those afternoon walks helped me to clear my mind and keep going.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Transportation models have grown in complexity by including more details on travel behaviour. However, some demographic and socioeconomic variables are still considered exogenous factors, introducing uncertainty and reducing the model's effectiveness in forecasting future scenarios. Since work trips are the most frequent trips in urban areas, understanding work-related attributes is relevant for travel demand modelling. Some attributes such as the place of residence, place of work, household income, auto ownership, and mode choice can directly or indirectly relate to the labour market.

According to Harmon and Miller (2020), the inclusion of labour market interactions within urban simulations has had little development despite the critical relationship between work and transportation systems in urban areas. In their paper, Harmon and Miller (2020) proposed an Agent-Based framework for simulating the demand and supply interactions of the labour market. This framework provided the first approach to a fully endogenous labour market simulation within a transportation-related model. Nevertheless, as discussed by Harmon (2013), much research still needs to be done to ensure the validity of results as the model evolves in time, to understand the interactions between agents, and to investigate the factors influencing the recruitment process within firms. In particular, He provides some evidence that the existent wage model could be underestimating salaries in the long run, which can be critical given the role of wages in the labour market.

Therefore, this thesis presents a model that predicts salary based on individual attributes of a worker using the principles of Bayesian inference. This model improves the prediction accuracy by accounting for the hierarchical structure of the data, which better simulates the variability in the salaries at both an aggregated and disaggregated level.

1.1 Outline

Although all sections in this document are structured sequentially, some can be optional according to the reader's knowledge of Bayesian inference. After this introduction, Chapter 2 presents an overview of economic theory and discusses the role of salaries and wages in the labour market interactions. Chapter 3 briefly introduces Bayesian inference, the framework for estimating the proposed model. Chapter 4 discusses the data sources and the hierarchy of labour data. Additionally, it presents the main results of the exploratory data analysis that guides the model specification. Then, Chapter 5 presents the details of the proposed salary model and the validation results with new data, followed by Chapter 6, which discusses the integration of this proposed model into the existing ILUTE framework. Finally, Chapter 7 compiles the principal results, and discusses the future work.

Chapter 2

Literature review

This chapter provides a systemic perspective on the labour markets supported by the literature. It starts with a brief discussion of the economic theory that governs labour markets, followed by an overview of different implementations of microsimulation models with an emphasis on the ILUTE framework. Finally, this chapter closes with a discussion of the gaps that guide the work in the subsequent chapters. Although the economic literature on this subject is extensive, this section is mainly based on the concepts and ideas stated by (Benjamin et al., 2021; Borjas, 2020; Kaufman and Hotchkiss, 2003; Smith, 2003)

2.1 The economic theory of labour markets

Labour markets are defined by the interaction of three principal agents[1]: workers, firms, and governments, each with their subgroups, objectives, and agendas (Benjamin et al., 2021). On one side, workers are trying to sell their labour at the highest price, the *supply side*, while on the other side, firms are trying to buy the labour at the lowest price, the *demand side*. The regulations imposed by governments constrain these interactions between workers and firms.

Different factors influence decisions made by individuals and companies and affect indi-

[1]Unions can be analyzed as an additional agent depending on local legislation, and the union's influence power, among others (Borjas, 2020)

vidual and group outcomes, such as earnings, type of employment, work hours and wages. Therefore, the primary objective of interactions within the labour market is to negotiate and agree on two key elements: wages and type of employment (Benjamin et al., 2021).

The wage is the element that facilitates the exchange between agents in the labour market but also allocates labour to the most efficient use in terms of occupations, industries, and regions within a market economy. This allocation encourages productivity improvements, job search and mobility, and investments in human capital through training and education (Benjamin et al., 2021). For firms, especially in sectors heavily dependent on the labour force, such as agriculture or manufacturing, wages are crucial in determining the prices of goods and services (Borjas, 2020).

In a perfect competitive model, wages are only determined by the interactions of supply and demand, but the assumptions of this model are rarely met. A more realistic model assumes that individuals are heterogeneous and that there are differentials in the wage determination. In this scenario, wages are determined by the market interaction given the individual characteristics of the worker and the job offered (Kaufman and Hotchkiss, 2003). Smith (1910) defines five principles that govern this differential in compensations: how pleasant or unpleasant is the job, how easy it is to learn the job, the stability of the job, the responsibility level, and the probability of success in that job. These principles are still applicable to most modern labour market configurations.

On the other hand, individuals and firms negotiate the type of employment, which refers to the characteristics of the job, such as the number of hours, flexibility, long-term stability, benefits, and associated risks. From the worker's perspective, the desired employment type is related to their values and preferences. Some individuals will be willing to work longer hours to earn more, while others prefer a balance between work and life. From the firm's perspective, the type of employment depends on the industry, the level of production, and the decisions within specific economic cycles. If the demand for a good is high and the economy is growing, a company would increase its labour force, while if the demand for the good drops, the firm will be forced to reduce its workforce.

Given the importance of wages and employment type in negotiation, the subsequent

sections review these concepts for both sides of the market, focusing on the short- and long-term decisions.

2.1.1 Labour supply

Microeconomic theory states that the worker's behaviour can be explained using the neoclassical model of consumption-leisure choice. In this theoretical model, individuals exchange their leisure time for the consumption of goods —represented by the income received from working. The combination of leisure and consumption provides some level of satisfaction, measured as *utility*, which is related to the values and preferences of each person. Hence, an individual entering the market aims to maximize their utility by choosing the best combination of leisure time and consumption.

Time is scarce, so individuals make decisions under a constrained scenario in which short-term choices are limited to the available daily time after sleep. In contrast, long-term decisions are primarily defined in extended periods and involve a more thoughtful process.

Short-term decisions

An individual's first and most important decision is whether to enter the labour market. According to Borjas (2020), every individual has a wage rate, known as *reservation wage*, that motivates them to work. This wage rate reflects how much an individual values their leisure time but also includes some information about other sources of income such as investments or inheritances —also known as *non-labour income*.

A person who highly values their leisure time or has a non-labour income covering their basic expenses could have a higher reservation wage. In comparison, someone indifferent to leisure time or without another source of income could have a lower reservation wage. Therefore, the decision to participate or not in the labour market will be given by comparing the market and reservation wage rates. If the market wage exceeds the individual's reservation wage, this person will be motivated to work; otherwise, this person will remain out of the labour market.

After deciding to participate in the labour market, individuals choose how many hours

to work. This choice is modelled by constructing labour supply curves using the same consumption-leisure model. In this model, any increment in the market wage above the reservation wage leads to working more hours until the individual's choice between leisure and consumption is optimal. However, this model assumes that people only allocate their time to leisure or work. Still, there are different activities related to household production that are not considered under this model (Borjas, 2020). To overcome this issue, the *household production function* allows the time allocation to be modelled based on the joint decision within the household unit. Benjamin et al. (2021) point out that this collective approach explains the recent changes in the Canadian workforce composition, in which more women are participating in the labour market, and the two-earner family groups are dominating the labour market structure.

In the household production function, the available time in a household is allocated based on the potential earnings of each individual, the dollar value of goods they want to consume, and the dollar value of the goods produced in the household[2]. Some findings in historical data have demonstrated that optimal time allocation for a household is given mainly by the potential earnings of the individuals, in which the person with higher wages tends to work more hours while the other individuals will increase their participation in household activities (Borjas, 2020). However, introducing different technologies in recent years has changed this optimal allocation by improving the production of household goods and services (Benjamin et al., 2021).

Long-term decisions

The leisure-consumption model assumes that individuals make decisions only focused on the present by assessing the factors prevailing in one specific period. However, McCurdy et al. (1980) demonstrated that individuals make long-term decisions based on the expected wage changes during the life cycle. In this model, any wage change has two effects: an increase or decrease in the hours dedicated to work.

In the first case, the *substitution effect*, any wage increase makes leisure time more ex-

[2]These goods are consumed at home and cannot be exchanged in any market. Some examples are childbearing, cooking, and cleaning.

pensive and motivates individuals to work longer. This effect is usually observed in people starting or at the mid-point of their professional careers. In the second case, the *income effect*, any increase in the non-labour income produces a rise in the utility level, giving the individuals the opportunity to work fewer hours. This effect is commonly observed among experienced individuals or people close to retirement. Besides the decisions based on the worker's life cycle, individuals also adjust their labour supply based on business cycles. According to Borjas (2020), during recession periods, some household members who are out of the labour force (*secondary workers*) can be motivated to enter the labour market if the *primary worker* loses their job.

Besides wages and earnings, some long-term decisions individuals make during their life cycle influence their choices related to work, such as childbearing, education, and retirement. These events might not be relevant to all individuals in the labour market but are closely related to individual preferences and the socioeconomic context. For childbearing, according to Angrist and Evans (1998), maternity reduces female labour supply with a lower effect among college-educated women or households where the male receives high wages. In contrast, they found minimal changes in the male labour supply in response to the birth of a child. They conclude that the decision to increase the family size is absorbed by the reduction of women's earnings or by purchasing child-care services, and the collateral effects are perceived both in the short and the long term based on the household income.

In the case of education, the human capital theory explains how individuals perceive the potential benefits of acquiring additional skills by forgoing current income in a similar approach to assessing an investment. Benjamin et al. (2021) argue that the motivation to enroll in a program is related to the wages the market is willing to pay. Under this perspective, firms are also eager to pay higher salaries and hire highly educated workers based on the assumption that these individuals will be more productive. However, the "ability at work" is not only obtained through education but can also be acquired through on-the-job training. This evidence suggests that both education level and experience can predict wages in the labour market.

Finally, the decision to retire is based on a combination of different factors such as wealth,

earnings, and the balance in the pension accounts. As described by Benjamin et al. (2021), the main factor is the wealth and earnings of the individual, in which the non-labour income plays a crucial role in determining the retirement age. They also find that other factors such as health conditions, the family and the nature of work can influence the decision based on the individual attributes. Some health conditions can motivate early retirement, especially in those sectors that demand physical activity from their workers —*blue-collar* jobs, while in other physically demanding jobs, retirement can be postponed —*white-collar* jobs. Also, changes in family composition and size in recent years could motivate people to remain longer in the labour market (Baker and Benjamin, 1999).

2.1.2 Labour demand

On the demand side of the labour market, firms are the agents that use labour to produce goods and services. In product markets, labour is considered a *derived demand* because it is one of the production factors. The relationship with product markets makes labour demand dependent on the prices of goods and services and the structure of these markets. That means that the decisions of firms operating in a competitive market will differ from those in a non-competitive market (such as monopsony or oligopoly). However, when firms enter the labour market, they aim to maximize profits and minimize costs regardless of the market structure (Benjamin et al., 2021).

Like the labour supply, firms' decisions vary between short- and long-term in relation to the main factors in the firm's production function: labour, capital, and technology (Borjas, 2020). In the short term, it is assumed that capital and technology remain fixed, and labour is the only factor firms can use to modify their production. While in the long term, all factors can vary.

Short-term decisions

As profit maximizers, firms will adjust their factors to achieve the lowest cost of production. While changes in some factors can take more time to be implemented —the construction of a new plant or the preparation for an IPO to raise equity capital from investors, labour is

the only factor that can be easily modified in the short term. Under this scenario, firms will adjust their labour demand based on the cost of labour (wages), their worker's productivity, the firm's production level, and the prices in the product market (Kaufman and Hotchkiss, 2003).

The *marginal productivity theory* evaluates these factors to model the firm's demand for labour under the assumption that the firm operates in a perfect competition market. Hence, a firm will increase its workforce by one worker if the productivity increment given by this additional worker is higher than the cost of hiring it —represented by the wage (Kaufman and Hotchkiss, 2003; Borjas, 2020). Given that the amount of capital is fixed, there is no substitution effect, and the labour demand will experience a diminishing marginal productivity and scale effect (Benjamin et al., 2021). That means there is an optimal point at which the number of workers maximizes the firm's profits, and after or before that point, the firm is in a non-optimal scenario. This optimal number of workers will be the firm's short-term labour demand.

However, quantifying the worker's productivity is not an easy task. Kaufman and Hotchkiss (2003) state that the difficulty of measuring individual productivity increases with the firm size, as more extensive production lines involve many workers in the production process.

Long-term decisions

As all factors can vary in the long run, labour can be substituted for other inputs such as capital. This substitution is based on the relative prices of both labour and capital. Therefore, the long-term decision to increase or decrease the labour force is subject to wages and the economic outlook that defines the cost of capital (Kaufman and Hotchkiss, 2003). Under a scenario where wage rates are increased, if the output remains the same, it will impact the quantity of labour demand because firms will use more capital instead of labour —a *substitution effect*. Likewise, if the amount of the output is reduced, the firm will further reduce their labour demand because it will be in a suboptimal scenario —*scale effect*.

Technology also plays a key role in the long-term demand for labour. As technological

breakthroughs increase productivity and efficiency, it also increases the demand for labour—scale effect. However, some occupations are replaced by innovations, making some skills obsolete —substitution effect.

2.1.3 Market interaction

Labour markets have unique characteristics that differ in structure from other product markets. Kaufman and Hotchkiss (2003) identifies some key differences, such as the item exchanged (labour) is embodied in the human being, the long-term nature of employment relationship, the heterogeneity of workers and jobs, and the diversity of markets. These differences impact the agents' decision-making process, the labour market interactions, and the market equilibrium.

In a perfect competitive single-market model, the equilibrium is reached when workers and firms agree upon the market wage rate. This market-clearing state assumes that all individuals and firms meet their objectives[3]. However, under an equilibrium state, the difference between what the firms are willing to pay and the wage rate is known as producer surplus. Likewise, the difference between what workers are willing to earn and the wage rate is known as worker surplus. The aggregate of producer and worker surplus are the economic gains for participating in the trade or entering the market —also known as *gains from trade* (Borjas, 2020).

Under this equilibrium state, the efficient allocation of workers to firms is the one that maximizes the gains of trade. As wages are the mechanism to allocate resources in the labour market, wages could be the primary driver of occupation choices and labour mobility between different markets, risk environments, and regions (Borjas, 2020). However, a static market equilibrium may be more a theoretical concept than a state reachable in the real world. According to Borjas (2020), many labour markets do not adjust so quickly to supply and demand shocks. He illustrates this concept using the market for new engineering graduates as an example, in which the market fluctuates between periods of excess demand for labour

[3]All individuals eager to work at the market wage are employed, and all firms eager to pay the market wage hire the labour force they required.

and periods of excess supply.

This fluctuation generates a cyclical trend in the entry wages for engineering graduates —also known as the *Cobweb model*. In a period of excess demand for engineers, entry wages tend to be higher as a reflection of the market interaction. Thus, a new generation of students enroll in engineering programs using the current information from the market. Nevertheless, this information failure could create a future excess of supply once the new generation graduates from engineering school and enters the labour market. Therefore, the idea that equilibrium is more oscillatory than static helps to represent many other markets characterized by systematic booms and busts (Benjamin et al., 2021; Borjas, 2020).

2.1.4 Job search and worker-job matching process

In the perfect competition model, firms and workers have all the information to enter the market. However, the information flow is asymmetrical because neither workers nor firms are fully informed when entering the market. This information asymmetry implies that agents invest time gathering information about the labour market before a match (Borjas, 2020).

During the search phase, workers can choose from a set of potential offers, each with some differences in the wage, while firms can choose from a set of applicants, each with some differentials in their skills and experiences. According to Benjamin et al. (2021), these differentials can be represented by the *wage offer distribution*, which reflects the probability distribution of the wage offers received by an individual. Although this distribution is limited to the job opportunities available to that individual at that time, it is a proxy measure that reflects the market value of the worker.

In the job-matching phase, the individual sets their *reservation wage* —sometimes called *asking wage*— and sequentially assesses different opportunities, comparing them with the reservation wage, until they reach one above their expectation. Although the individual can wait to meet their expectations, there is an opportunity cost for not taking an offer below the reservation wage. This opportunity cost adjusts the reservation wage based on their individual preferences. Some individuals will have enough patience and savings to wait for the expected offer, while others will adjust quickly their reservation wage to match the

wage offer distribution (Benjamin et al., 2021; Borjas, 2020). Nevertheless, the adjustment of the reservation wage can also be performed by comparing it with the current market value distribution —represented by the wage offer distribution. If the reservation wage is outside the distribution, the individuals will adjust their expectations accordingly to increase the likelihood of being hired.

2.1.5 Wage definition

On the demand side, salaries can be defined by a *Hedonic model* that explains the wage based on the worker's perceived job risk (Borjas, 2020). Hence, riskier jobs have higher wages to incentivize people to accept them. However, the hedonic model can be extended beyond the definition of job risk by including other job attributes such as skill level, location, amenities, and long-term stability. Equation (2.1) presents the general form of this model in which w_i is the wage of worker i, ρ_i is the probability of injury, β_j is the parameter for the j^{th} job characteristic, and V_{ij} is the j^{th} job characteristic of the worker (Borjas, 2020).

$$w_i = \alpha + \sum_{j=1}^{J} \beta_j V_{ij} + \rho_i \tag{2.1}$$

On the supply side, human capital theory and life-cycle labour supply models have also used linear models to estimate wages as a function of several individual attributes. In human capital theory, wages reflect the accumulation of capital such as education and job experience, while in the life-cycle labour supply approach, wages are defined by some individual attributes relative to each industry, also known as *fixed factors* (Borjas, 2020; Sullivan, 2010).

2.2 Practical approaches for modelling labour markets

This section explores some implementations of labour market microsimulation models that manage earnings, transitions, and events as endogenous components. As the work of this thesis is expected to be used within the transportation and urban planning setting, the first

part reviews the urban microsimulation framework ILUTE[4]. The second part presents an overview of some microsimulation frameworks used for modelling labour markets in other contexts, such as demographic analysis and pension policy testing, to compare other approaches outside the transportation context.

2.2.1 ILUTE framework

The Integrated Land-Use, Transportation, and Environment framework (ILUTE) is an urban microsimulation model that simulates the activities of individual agents as they evolve in a specified period. This framework has been tested using data from the Greater Toronto Area (Salvini and Miller, 2005). In this modelling system, census population data are synthesized to produce agents that represent the population of the GTA and simulate different high-level decisions (residential location, firm location, auto ownership) and low-level decisions (activity scheduling, trip mode selection). The interactions between different agents are represented as market transactions, and the outcomes are the inputs to other sub-markets.

The ILUTE implementation consists of different sub-models that simulate interactions or processes in urban areas, as shown in Figure 2.1, in which the labour market was a sub-model "Under development" by the time that document was released.

Posterior works of Hain (2010) and Harmon (2013) proposed a procedure to simulate labour markets within the ILUTE framework. Hain (2010) prepared a series of linear regression models to predict wages based on some attributes and characteristics of the worker. In this work, he explores different predictors of wages, distinguishing between full-time, part-time, annual, and hourly salaries following the hedonic econometric approach reviewed in Section 2.1.5. The same approach was used to estimate the hours of work based on the attributes of the worker, but the linear regression models were not statistically significant.

In addition, Hain estimated a transition model using three binary logit models that

[4]To the author's best knowledge, most of the land use-transportation interaction models treat labour markets and employment exogenously, and ILUTE is one of the first to include it as an endogenous component. Harmon (2013) presented a comprehensive literature review about the approaches to include labour market on different land use-transportation models but most of them are focused on just one side of the labour market, the demand or supply. By the time this thesis is written, most of these findings remain the same.

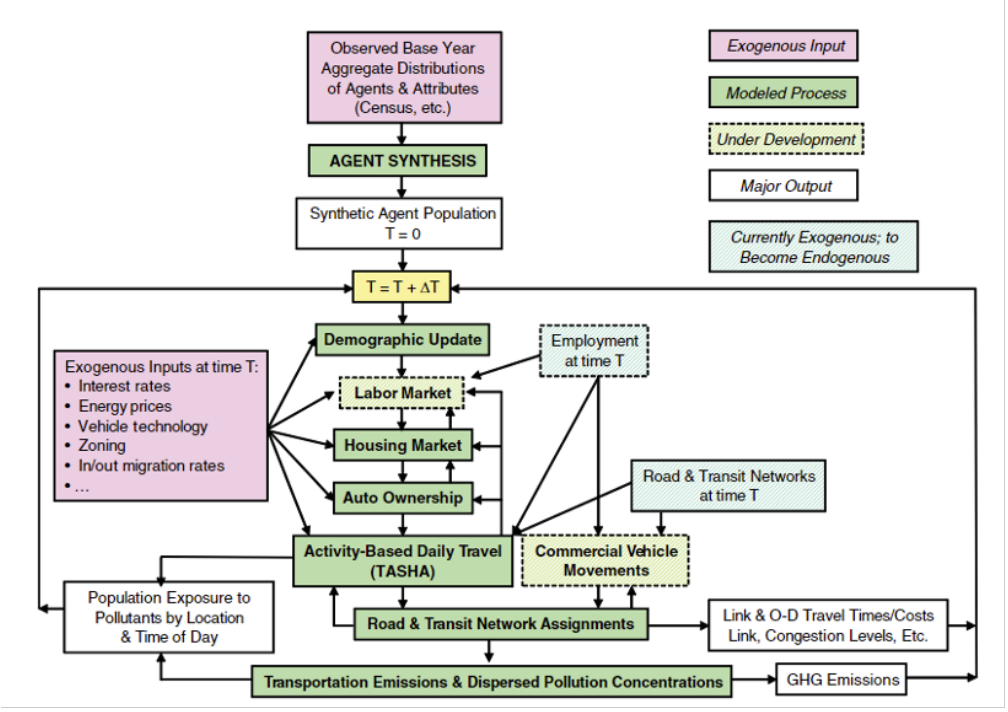

Figure 2.1: Flowchart of ILUTE processes (Miller and Vaughan, 2021)

calculate the probability of a person transitioning between being employed to unemployed, unemployed to out of the labour force, and out of the labour force to unemployed. Hain recognized the need for a fully endogenous microsimulation model of the labour market that treats both the demand and supply sides as agents.

Based on Hain's work, Harmon (2013) built an agent-based model using the wage and transition models. He proposed a simplified firmographic model representing the endogenous creation and deletion of jobs in the labour market. He also used a series of linear regression models based on GDP and unemployment rate to predict each period's added or deleted jobs per sector. Moreover, he estimated a series of cumulative distribution functions to assign different attributes to the worker and job class, such as industry, occupation, job type (full or part-time), firm size, and location.

Using these models, Harmon proposed a job-matching algorithm in which a random pool

of candidates is evaluated, and the worker with higher utility accepts the offer (Harmon and Miller, 2018, 2020). This job-search and matching process is summarized in Figure 2.2. All models are updated for each timestep (defined in ILUTE as a year), simulating the labour market's long-term dynamic as agents and economic conditions evolve over time.

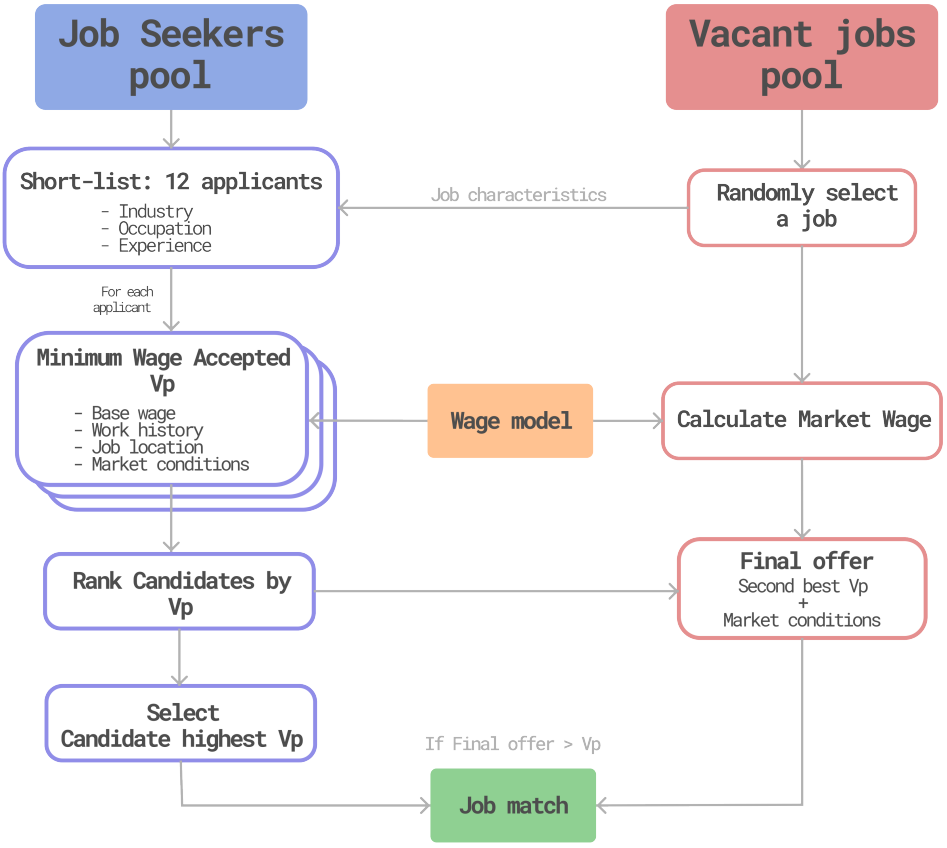

Figure 2.2: Job search and matching process in ILUTE (Adapted from Harmon (2013))

The simulation results have been compared with actual data from 1987 to 2006, demonstrating an accurate estimation of employment counts by industry and occupation. However, as Harmon (2013) pointed out, more research is needed to fully represent the internal interactions in the labour market. Both Hain and Harmon propose that this should be accomplished by generating a fully endogenous firmographic model that simulates the birth and death of firms while providing the evolution of jobs (creation and deletion) during these business

cycles.

Moreover, according to Harmon (2013), the wage model underestimates wages in relation to prior experience and tenure levels, directly impacting the matching algorithm's long-term stability. As the models were estimated using single-year datasets, Harmon points out the need to validate the temporal stability of the parameters for the job demand and supply models.

2.2.2 Other dynamic microsimulation frameworks

The following models are examples outside the transportation planning setting that use a microsimulation approach to assess the long-term impact of policy changes in the pension and fiscal systems. Although some models are being updated or replaced by new versions, this review provides evidence about the possible challenges and opportunities of implementing these microsimulation systems.

DYNACAN and Lifepaths (Canada)

Statistics Canada has a long history of experimenting with dynamic microsimulation models to estimate and project some sociodemographic characteristics relevant to Canadian society, such as aging, pension plans, and health. This effort has produced several models, from DYNACAN in the early 1990s to the LifePaths in the early 2010s. DYNACAN and LifePaths are discontinued because Statistics Canada is developing a new dynamic socio-economic microsimulation tool[5].

The DYNACAN model is a dynamic microsimulation model used to estimate the impacts of policy changes in demographics, earnings and other characteristics related to the Canadian Pension Plan -CPP (StatisticsCanada, 1998). For the base year, DYNACAN uses over 212,000 agents representing a sample of the Canadian population for the 1971 census (starting database). The model consists of three modules: Part A assembles and prepares the data into a single hierarchical database for the following modules, Part B simulates the

[5]This information was retrieved from: `https://www.statcan.gc.ca/en/microsimulation/index`, in February 2023

demographic events and employment earnings over time, and
Part C calculates the CPP contributions and benefits for each agent.

Regarding the labour market, DYNACAN models several events directly related to individuals' participation in the labour force. The events and the explanatory variables are summarized in Table 2.1.

Event	**Explanatory Variables**
Education	Gender, age, parents' education, living arrangements, house-ownership status, marital status, children
Employment status (worked less/more than 48 weeks)	Gender, age, marital status, children, disability status, retirement status, and working status in previous years
Female exclusion from the labour force	Age (>21), gender (female), marital status, children, labour force participation in previous years
Annual weeks worked (0, 1-47, 48 or more)	Age, gender, marital status, tenure in current job, education level, children, children's age, unemployment rate
Weakly wage	Age, gender, marital status, children, education level, unemployment rate, previous year earnings, employment status, tenure in current job
Income from employment	Product of wage time weeks worked
Retirement age	Age, disability status, gender, CPP contributions

Table 2.1: Labour market-related events on DYNAMO (Adaptation from Anderson (1997)

Although there is no official documentation about the procedures in this model, most of the events combine rule-based models and linear or logistic regressions that capture the agent's behaviour (Anderson, 1997). All simulations are performed using the Monte Carlo technique. Since the model is not a market-clearing model, there is no interaction between firms and workers and no feedback from the pension system provision, labour market, savings or demographic modules (Anderson, 1997).

DYNACAN was replaced by LifePaths, which was intended to be a multipurpose microsimulation model that simulates a large sample of individuals in much more detail than the DYNACAN. LifePaths uses a variety of Canadian microdata, which makes it suitable for analyzing government policies of a longitudinal nature. It supports different dimensional analyses such as cross-sectional, over individual lives, between cohorts and over generations (StatisticsCanada, 2013). LifePaths simulate different life events through probabilistic functions, and events are modelled as random processes in which the occurrence of an event is conditional to the previous history plus a stochastic component.

Similarly to DYNACAN, LifePaths simulates different labour market-related events based on the sociodemographic characteristics of each agent. Table 2.2 summarizes these events and their explanatory variables.

Event	**Explanatory Variables**
Student Employment	Gender, province of residence, status in the parental home, variables representing seasonal patterns
Employment status	Gender, age, province of residence, education level, children, marital status, spouse's employment status
Retirement (if age>60)	Age, gender, province of residence, family characteristics, and education level.
Hours worked per week	Age, gender, education level, province of residence, family characteristics, student status, job tenure
Maternity leave	Age, marital status, education level, employment status, tenure in current job
Weekly earnings	Age, gender, education level, field of study, province of residence, and immigration status

Table 2.2: Labour market-related events on LifePaths Adaptation from (StatisticsCanada, 2013).

Lifepath's official documentation does not provide detailed information about the procedures and models. It only states that the behavioural component combines rule-based models with logistic and linear regressions.

APPSIM (Australia)

The Australian Population and Policy Simulation Model (APPSIM) is a microsimulation model developed by the National Centre for Social and Economic Modelling at the University of Canberra to evaluate the impact of future social and fiscal policies until 2050. According to Keegan (2007) APPSIM uses a 1% sample of the 2001 census as the initial population. The primary individual characteristics are obtained from the census, but other microdata sources are used to impute additional characteristics such as earnings and employment history. Using a series of multinomial logit models, the model uses these data sources to estimate the transition probabilities for different states. The model updates the agent's states annually using Monte Carlo simulation.

APPSIM is structured as a modular framework in which agent states are updated annually. The modelled states range from demographic events such as births, leaving parents' home, marriage, and education to more specific ones such as the states in the labour market or health conditions. Regarding the labour market module, APPSIM classifies states in labour force status, employment dedication, and employment type. One essential process in APPSIM is the "Alignment" of simulation results with historical data. This process ensures that the microsimulation is aligned with aggregated results in macro projections and that the model results are stable in the long term.

ELSI (Finland)

The ELSI framework is a microsimulation model created by the Finnish Centre of Pensions to analyze employment trajectories in the context of the pension system. This model is an individual-level simulation alternative to the semi-aggregated model (LTP) used by the Finland Ministry of Finance. One of the advantages of ELSI is the ability to model immediate and long-term effects of policy reforms related to the labour market and pension system (Tikanmäki and Lappo, 2020).

ELSI uses a different approach than DYNACAN, LifePaths, and APPSIM because the transition probabilities are not obtained using behavioural equations (logistic regressions or multinomial logit models). Instead, ELSI calculates the transition between states using

probability functions following a random Markov process in which the future state only depends on the current state.

For the labour market model, ELSI defines 21 states in six categories: employed, unemployed, outside the labour force, sick, retired, and deceased. Given that the purpose of this model is to evaluate pension-related policies, the granularity of states is related to the relationship with the pension system in that state.

One key component of the ELSI model is that the earnings model calculates the labour and pension-related earnings separately, which generates a feedback loop with the labour market transition module. This representation introduces the effect of non-labour income in the transition decisions in the labour market. The annual salary is estimated using a hedonic regression based on some individual's characteristics and adds a random component to simulate the variability of wages between similar individuals. Tikanmäki and Lappo (2020) use an asymmetric Laplace distribution for modelling this random component.

2.3 Conclusions of the literature review and current gaps

The literature review presented in this chapter provides evidence that labour markets possess structural differences from other markets in the economy because the product, labour, is attached to individuals. This attachment increases the complexity of individual decisions in the short and long term because the possible outcomes can significantly impact other parts of life. Moreover, much of the information the agents share in labour markets is incomplete or asymmetrical, affecting market efficiency and producing sub-optimal equilibrium.

In this context, microeconomic theory is helpful to explain the rationality behind decisions in the labour market. However, the assumptions of these models do not fit with most of the actual conditions of labour markets. The field of econometrics uses observed data to evidence the theoretical structure of labour markets in an aggregated or semi-aggregated way. However, these models can be extended into a microsimulation approach that captures the heterogeneity of the agents and better represents the interaction between firms and workers.

Although most examples of microsimulation models in the labour market have been developed with a unique purpose, a new generation of dynamic models is developed with a multipurpose perspective. ILUTE follows this idea by modelling different systems within the urban context with the idea that the world operates as a system of systems.

The results obtained by Harmon using the Agent-Based implementation of the Labour Market module in ILUTE demonstrate that the microsimulation approach is a good representation of the market at an aggregate and semi-aggregated level. However, additional efforts should be focused on improving the models that govern the behavioural decisions within the module.

In particular, Harmon points out that the existent wage model underestimates salaries when comparing the intercepts of the full-time with the part-time models. One possible explanation for this issue is the quality of the values of worked hours reported in SLID. As this survey was performed once every year, respondents could not recall this number correctly. Therefore, a possible approach is to use annual salary instead of hourly wage. This variable, also reported in SLID, could be more reliable and produce better results.

Moreover, the deterministic nature of Hain's model could be an additional source of the log-run instability of the model. As the linear model represents the average salary for an individual with specific attributes, all individuals sharing those attributes will have the same predicted salary. From the simulation perspective, this approach is missing an essential component of reality, in which randomness can explain some differences in wages between individuals. Therefore, a model that includes this random component could improve the stability and behavioural representation of the agent-based model.

This deterministic outcome could also affect the current implementation of the job-matching process. The current price-taker approach in which workers accept or decline a job offer does not consider that they assess the market by applying to different job offers and select the most optimal. Although Harmon's model adjusts the *MinimumWageAccepted* by location and number of attempts, the wage model produces the same point estimate for all individuals with the same attributes, which translates into all job offers in one specific bucket being equally perceived by the workers. In reality, the wage differentials between

offers make some more appealing than others, and the trade-off between the benefits and costs[6] of each option is the decisive factor. Hence, including the random component in the salary representation in the ABM module could improve the job-matching process.

Given these challenges and the importance of wages in the interaction between workers and firms in the labour market, this thesis proposes an alternative salary model that implements the missing random component and introduces some changes in the current structure of the Agent-Based model to represent the job-searching and matching process better. The proposed model uses the Bayesian inference approach to model the uncertainty in the prediction and produce probability distributions instead of point estimates. The details of this approach are covered in the following chapters.

[6]Such as transportation and travel time costs

Chapter 3

Bayesian inference: From point estimates to probability distributions

This section presents an overview of the concepts in Bayesian inference used for the model estimation in the following chapters. It starts with a brief comparison between the Frequentist and Bayesian approaches. Then, it presents the theoretical foundations of Bayesian statistics and how these principles are transformed into a practical framework for parameter inference and modelling in the real world.

This chapter is based on the ideas discussed by (Lambert, 2018), (McElreath, 2016), and (Gelman et al., 2013) but is not intended to be an extensive review of the concepts. For further details, these sources are a good starting point.

3.1 The Bayesian paradigm: Frequentist vs Bayesian

Statistical inference has two "Schools of thought": the Frequentist (or Classical) and the Bayesian approach. For frequentists, the data are assumed to be the result of an infinite number of repeated experiments with the same characteristics. Then, the data are randomly sampled from a fixed and defined population, and any source of variation comes from that sampling process. Under this perspective, model parameters are assumed to be fixed but

unknown values related to the population of interest, and the objective of inference is to calculate the best point estimate of the true value of the parameters given a data sample.

In contrast, Bayesian statistics assume that data are observed and fixed quantities, and the source of variation comes from the uncertainty over the parameters. From this perspective, parameters are probabilistic values, and the objective of inference is to estimate the probability distribution of the model's parameters. Then, we use the data as evidence to update any prior belief about the underlying process.

The debate about which approach is the best is exciting but long and almost philosophical; therefore, it is outside the scope of this document. However, as (McElreath, 2016) points out, the Frequentist approach could make sense if the process of interest can be replicated multiple times, as in the case of many natural sciences (i.e., multiple controlled experiments in a laboratory). In contexts where the data collection can only be performed once, such as in many social sciences (i.e., democratic elections, population census), a Bayesian approach could be more aligned with the data nature.

A quick overview of the Bayesian thinking framework is presented in Equation (3.1) (Lambert, 2018). The procedure begins with establishing prior beliefs about the process under analysis. Then, evidence (data) is gathered to update the prior beliefs using the model. The update is known as the posterior belief that includes the knowledge from both the priors and the data.

$$\text{prior} + \text{data} \xrightarrow{model} \text{posterior} \tag{3.1}$$

3.2 The basics of Bayesian inference

As the aim of statistical inference is to estimate the parameters of a model that recreates the process of interest, one of the objectives is to calculate the probability of getting the true parameters given the observed data or $P(\theta|data)$. However, as the true parameters are unknown, only the likelihood of the data being generated by the model parameters $P(data|\theta)$ can be calculated.

According toLambert (2018), both Frequentist and Bayesian inference aim to go from $P(data|\theta)$ to $P(\theta|data)$. While in the frequentist perspective, this transformation is not correctly carried out[1], the Bayesian perspective applies the Bayes rule –Equation (3.2)– to properly transform the likelihood into a probability $P(data|\theta) \rightarrow P(\theta|data)$. The details of this transformation are reviewed in the following sections.

$$P(\theta|data) = \frac{P(data|\theta) \cdot P(\theta)}{P(data)} \tag{3.2}$$

posterior · likelihood · prior · evidence

3.2.1 Likelihood

One of the most important tasks when performing statistical inference is the choice of the model and its parameters. As the parameters are unknown to the modeller, the inference task is to estimate their value. One approach is to test different parameters on the chosen model and calculate the probability of getting the observed data for each set of values. The function that describes these probabilities corresponds to the *likelihood function*. However, as this function is not a probability distribution, Bayesian statistics use Bayes theorem to convert this likelihood into a proper probability distribution.

When defining the model, intrinsically, the data generation process is considered. Therefore, the likelihood includes the evidence and all the assumptions about the process. The higher the likelihood[2], the better the estimation of our true parameter distributions. Given the directionality of the Bayesian thinking process $prior + data \rightarrow posterior$, the model def-

[1] The estimation is done by selecting the parameters that maximize the likelihood of obtaining our observed data.

[2] In practice, the likelihood is transformed into the log-likelihood to ease the computation. This transformation also modifies the optimization process into a minimization. However, both approaches yield the same results.

inition is one of the most important factors in Bayesian inference. If the model choice does not represent the data generation process, no matter the prior selection, the inference process will fail to produce parameters that represent the reality.

3.2.2 Prior

Also known as prior belief, it is the distribution or group of distributions[3] representing previous knowledge of the process under study. This distribution encodes the existing knowledge and the modeller's uncertainty of the model parameters.

In some cases, the problem is widely studied, and there is enough evidence about the possible values of the model parameters. In other cases, there is no previous evidence about the process. In the former case, the modeller is more confident about the expected value of the model parameter and can approximate the shape and location of the probability distribution for the model parameters. In the latter case, the modeller is more uncertain about the possible parameter values, and a conservative approach is the best choice. Then, the prior choice represents the modeller's current knowledge of the process under study.

The prior selection is subjective as the current knowledge state can vary from one modeller to another. McElreath (2016) argues that all modelling approaches, Frequentist and Bayesian, involve a certain degree of subjectivity. However, the Bayesian approach makes it explicit and open to scrutiny. If the current belief contradicts the evidence, the modeller can use the Bayes theorem to include this new evidence and update their prior belief.

Prior distributions are classified by the knowledge state, as shown in Figure 3.1(a). *Highly informative priors* are concentrated around a value, and *non-informative priors* are widely distributed and assume almost equal probabilities given the uncertainty level. Likewise, the shape and distribution type choices are crucial for correct prior selection. Figure 3.1(b) shows some highly informative prior alternatives for modelling continuous and positive variables. When modelling processes with random components or in models that include linear predictors (e.g., linear regression, logistic regression), the parameters are usually represented by the normally distributed prior. This prior provides more information than the uniform

[3]There is one prior for each model parameter.

but is more flexible to be updated in the process. These distributions are also known as *weakly-informative priors*.

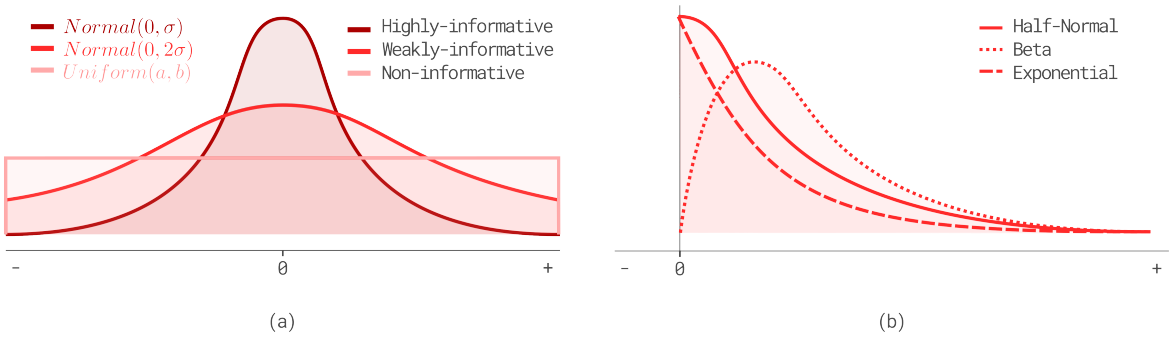

Figure 3.1: Types of priors

3.2.3 Evidence

Also known as *marginal likelihood*, it represents the probability of observing the data considering all possible parameter values. It is called marginal likelihood because it corresponds to the marginal probability calculated as the integral of the joint probability $P(data|\theta)$ across all θ's, as represented in Equation (3.3)(Lambert, 2018).

$$\begin{aligned} P(data) &= \int_{all\theta} P(data|\theta) \cdot P(\theta) d\theta \\ &= \int_{all\theta} P(data, \theta) d\theta \end{aligned} \tag{3.3}$$

Although Equation (3.3) is the formal definition, the marginal likelihood has an additional interpretation. Given that the likelihood function is not a probability distribution, it is expected that the multiplication of the likelihood by the prior is not either. Therefore, as the posterior needs to be a probability distribution, the marginal likelihood is the normalization factor that scales the numerator, so the area under the posterior distribution sums up to

1. Under this interpretation, the numerator in the Bayes theorem provides the shape of the posterior, while the Marginal Likelihood scales it out[4].

The calculation of Equation (3.3) can be trivial for simple models, but it becomes intractable for some real problems that include several parameters. This is one of the reasons why most of the Bayesian inference must be addressed using sampling methods instead of exact calculation. This will be reviewed in more detail in Section 3.3

3.2.4 Posterior

Finally, applying the Bayes theorem to update the priors based on the evidence results in the posterior distribution. The posterior represents the updated probability distribution of our model's parameters and consolidates the knowledge of our system that was extracted from the data and the priors.

While the Frequentist approach produces a point estimate for each parameter in the model, the Bayesian approach produces a probability distribution for each parameter, representing the uncertainty over the estimated model parameters. Then, point estimates such as mean, median, mode, and interval estimates can be calculated from this distribution.

Although the point estimates produced by Frequentist and Bayesian approaches can be similar, there is a difference in the interval estimates produced by both approaches. In the Frequentist approach, the confidence interval (CI) is an interval of plausible values of a population parameter with a certain level of confidence constructed over many repeated experiments (Devore, 2016). For instance, a 95% confidence level implies that if an experiment is repeated many times, 95% of the samples will produce a CI that contains the true population parameter. Therefore, the confidence interval is a measure of the uncertainty over the interval, not over the parameter estimation (Lambert, 2018).

In the Bayesian counterpart, the credible interval is a range that contains the true value of a population parameter with a particular probability. For instance, a 95% credible interval means that the given range contains the true value with a probability of 95%. Conversely

[4]This abstraction is key to introduce the sampling methods reviewed in Section 3.3

to the Frequentist counterpart, the credible interval is a measure of the uncertainty over the estimation parameter. This interpretation of interval estimates seems more intuitive than the Frequentist confidence interval because it provides a probability over the parameters of interest.

3.3 Applying Bayes theorem: Markov Chain Monte Carlo sampling

As discussed in the last section, the calculation of the denominator in the Bayes theorem complicates the application of Bayesian inference to real problems. The solution is to abandon the idea of exact calculation and use alternative methods to estimate the posterior distribution. One approach is to sample from the distribution to build an approximation; however, the posterior is still unknown until we can solve the denominator.

Coming back to the idea that the numerator of the Bayes theorem provides the shape of our posterior and the denominator is a scaling factor, we can rewrite the Bayes theorem as:

$$\begin{aligned} P(\theta|data) &= \frac{P(data|\theta) \cdot P(\theta)}{P(data)} \\ P(\theta|data) &\propto P(data|\theta) \cdot P(\theta) \end{aligned} \tag{3.4}$$

In Equation (3.4), the posterior is proportional to the multiplication of the likelihood function and the prior. Given that both terms in the numerator are known[5], it is possible to sample from this function even if the denominator is unknown (it is not normalized), as shown in Figure 3.2.

[5]Priors are defined by the modeller and likelihood is calculated based on the model choice and the observed data

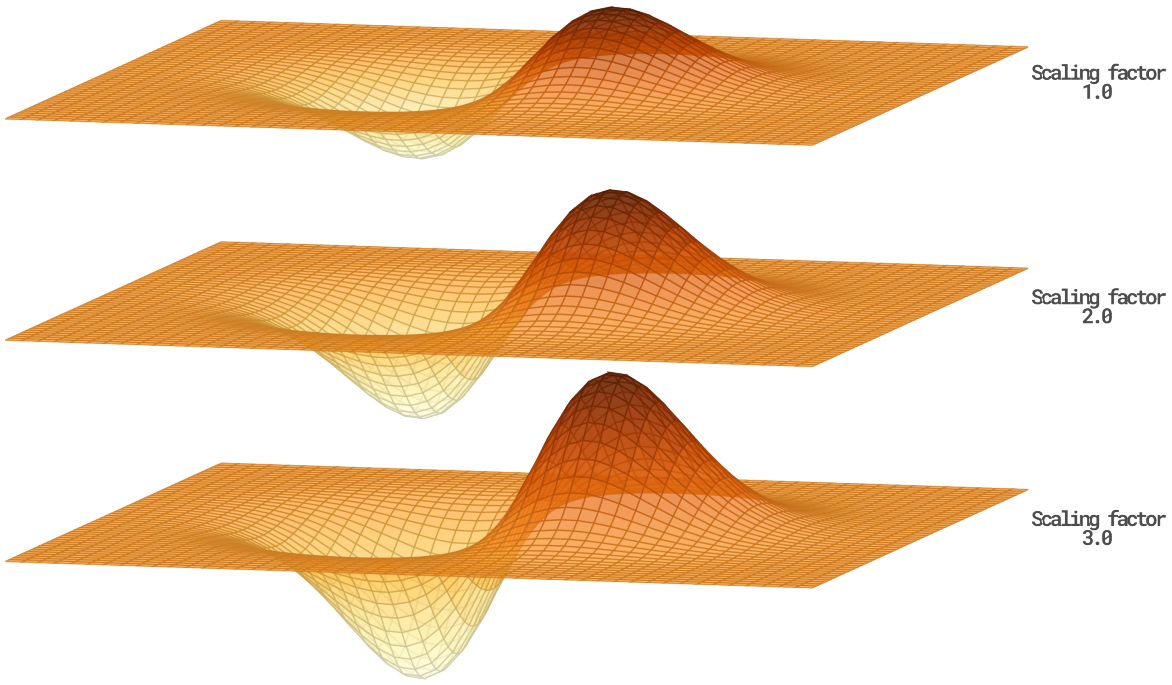

Figure 3.2: Effect of scaling factor in the posterior space

This sampling process uses a method known as Markov Chain Monte Carlo (MCMC), which allows the approximation of a probability distribution by obtaining a sequence of random samples. The following section briefly introduces one algorithm within the MCMC family called Hamiltonian Monte Carlo (HMC), which is used in the development of this thesis. For further details of this algorithm, please review (Lambert, 2018) and (Neal, 2012).

3.3.1 Sampling from posterior: Hamiltonian Monte Carlo

As the objective of the sampling process is to approximate the shape of the posterior distribution in Equation (3.4) (also known as *posterior space*), different approaches such as rejection sampling, Gibbs sampling, or a simple Random Walk sampling can be used. However, these

techniques are inefficient because they do not consider the shape of the distribution[6].

The *Hamiltonian Monte Carlo* method (HMC), a special case within the family of Markov Chain Monte Carlo (MCMC) algorithms, explores the posterior space more efficiently by using an analogy of a physical system in which a frictionless particle is moving around this space. In this physical system, two forces interact with the particle: gravity and the initial random momentum[7]. As the system is frictionless, the particle's position can be obtained at any time based on the previous position and the momentum applied to the particle (Gelman et al., 2013; Lambert, 2018; McElreath, 2016; Neal, 2012).

To begin with, the posterior space is transformed into the negative log of the posterior $-log[(P(data|\theta)) \cdot P(\theta)]$, also known as NLP space. This transformation converts all the peaks in the posterior space (zones with higher probability density) into valleys and the valleys in the posterior space (zones with lower probability density) into peaks. The idea behind this transformation is that the particle will visit more often the valleys of the NLP space because of the gravity effect. Therefore, more samples are taken from the zones with high probability density in the posterior space. Figure 3.3 shows a representation of the posterior and NLP space. It is important to highlight that the dimension of the NLP space is equal to the number of parameters in our model. So, the coordinates in this space are the model parameters.

[6] An efficient sampler produces more samples from areas in which the posterior peaks (high probability density) and fewer samples from the rest of the distribution.

[7] This is the initial random impulse the particle receives that allows it to explore the posterior space.

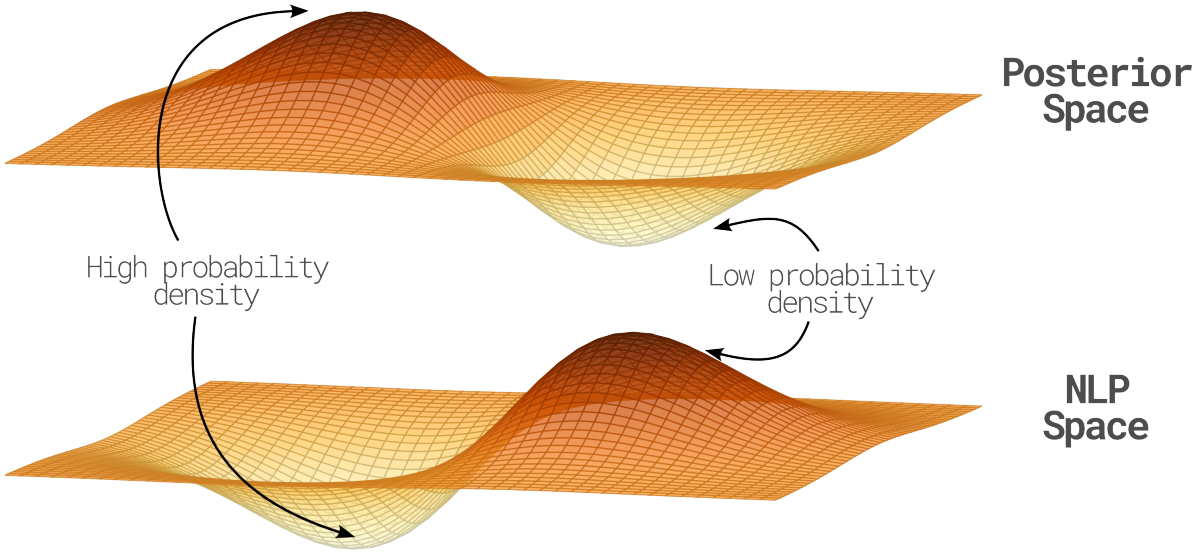

Figure 3.3: Sampling from the posterior distribution and the NLP space

After this transformation, the initial position of the particle θ_o is randomly initialized[8] and the following iterative process starts:

- First, an initial momentum is applied to the particle to start the process. This momentum is sampled from a multivariate normal distribution $m \sim \mathcal{N}(\mu, \sum)$. Usually, this distribution is centred at zero with a diagonal covariance matrix.
- With this initial push, the particle will move through the sample space for a predefined amount of time T[9]. Then the new position θ_1 and the momentum at that position m_1 are saved[10].

[8] θ_o is a vector of coordinates in the posterior space

[9] This parameter is associated with the convergence of the algorithm. For instance, a smaller T will better explore the NLP space but it will take longer to converge

[10] In this part, the algorithm called *leapfrog* solves the path of the particle moving over the NLP space using the Hamiltonian dynamics. For further explanation of this algorithm, please review (Neal, 2012)

- With the initial and final position, the probability $r = \frac{P(X|\theta_1)\cdot P(\theta_1)}{P(X|\theta_o)\cdot P(\theta_o)} \times \frac{q(m_1)}{q(m_o)}$ is calculated. In this expression:
  - The first fraction is the ratio of the un-normalized posterior space at the initial and final position[11].
  - The second fraction is the ratio between the probability density function used to sample the momentum, evaluated at the initial and final position. This fraction allows the sampler to visit zones of lower probability (peaks in the NLP space) that otherwise would be rarely sampled because gravity biases our sampling to zones of higher probability.
- A random number $u \sim Uniform(0, 1)$ is generated. If $r > u$, position θ_1 is accepted, and the particle will start the next iteration from that position; otherwise, the particle will return to θ_o.
- Each accepted position θ_i is saved into a list. The sequence of θ_i samples is known as the sample chain. It is important to highlight that the accepted position θ_i corresponds to a set of samples of all model parameters because each coordinate in the posterior space corresponds to a parameter in the model. Figure 3.4 illustrates the sample-gathering process and details how the posterior distribution for each model parameter is built as the HMC algorithm explores the posterior space.

[11]The denominator in the Bayes theorem would have been cancelled in this expression because is the same for all places in the posterior space. Therefore, this is another explanation of why the denominator is not needed to sample from the posterior distribution.

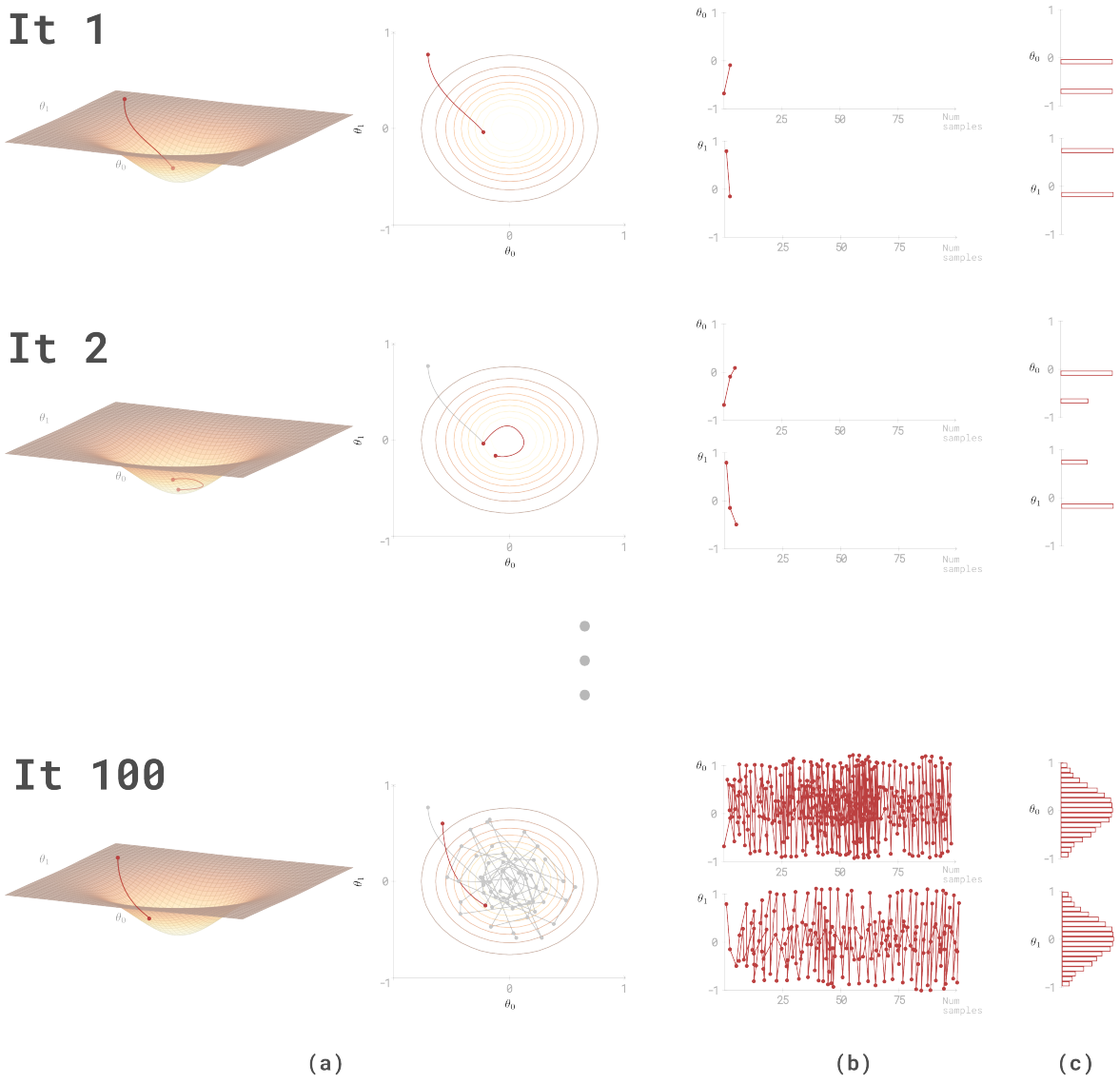

Figure 3.4: Sampling process using the Hamiltonian Monte Carlo Algorithm

This iterative process is done in parallel for multiple chains to evaluate the sampling convergence and ensure that the posterior space is thoroughly sampled by starting in multiple

random positions. The idea behind the multiple chains is that the posterior space will be better explored if each chain mixes well with other chains and does not get stuck in a single place, as shown in Figure 3.5.

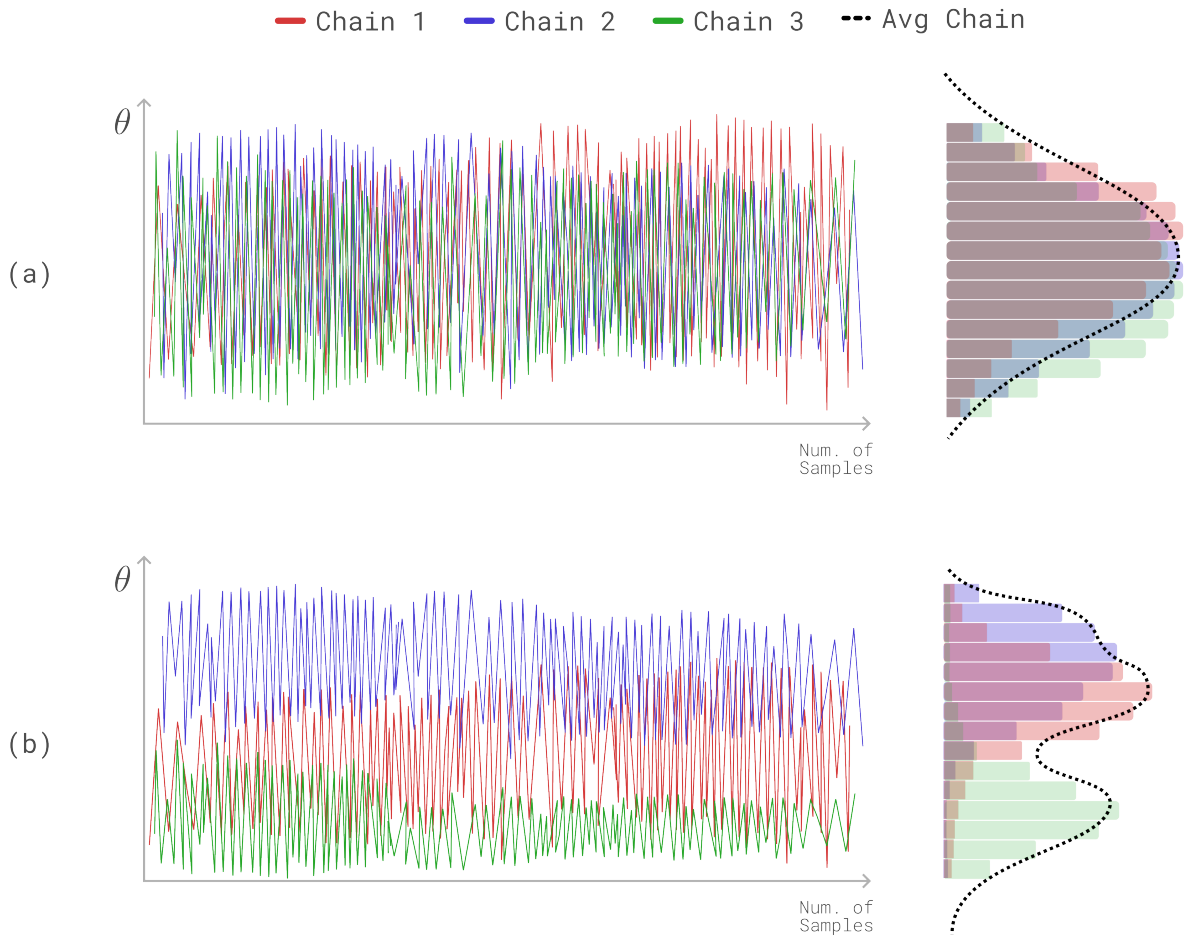

Figure 3.5: Sampling using multiple chain exploration

A common practice is to discard several iterations at the beginning of the process. These iterations are known as the warmup of the model and allow the chain to be stable before collecting the samples. Then, the sampling process is repeated multiple times until conver-

gence is guaranteed. Convergence is measured by calculating the within- and between-chain variation, as proposed by Gelman and Rubin (1992),using the statistic $\hat{R}$ in Equation (3.5).

$$\hat{R} = \sqrt{\frac{W + \frac{1}{n}(B - W)}{W}} \tag{3.5}$$

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2 \qquad ;\text{with} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta_j})^2 \tag{3.6}$$

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{\theta_j} - \bar{\theta})^2 \tag{3.7}$$

Where, W is the within-chain variance, s_j^2 is the estimator for the sample variance for j-th chain, n is the number of samples, m is the number of chains, and B is the between-chain variance. A common rule of convergence is that the statistic $\hat{R}$ should be below 1.01 for each model parameter.

For the analysis performed in this document, a variant of the Hamiltonian Monte Carlo algorithm is used to increase the sampling efficiency. This variant is known as the No U-Turn Sampler (NUTS) and follows the same principles as HMC but dynamically adapts T to increase the sample's acceptance rate.

3.3.2 Posterior predictive distribution

The last section focuses on the use of MCMC methods (HMC in particular) to estimate the posterior distribution of model parameters, which can be used to draw conclusions or perform hypothesis testing. However, the real power of the model is when it is used to predict values in the light of new data. After HMC or any other MCMC algorithm is used to estimate the parameter posterior distributions, these posteriors can be used for prediction.

The prediction in the Bayesian inference is performed using the following steps, also illustrated in Figure 3.6. Although this example corresponds to a univariate linear regression

model with one intercept and one slope as model parameters, these steps can be applied in the same way to the multivariate case or any non-linear model:

- Samples from the posterior distribution of the intercept and the slope are collected.
- Each set of parameter samples (intercept-slope) is used to calculate a regression line that produces a new target value.
- These target values are known as the *Posterior Predictive distribution*, which is the posterior distribution of the target variable[12].

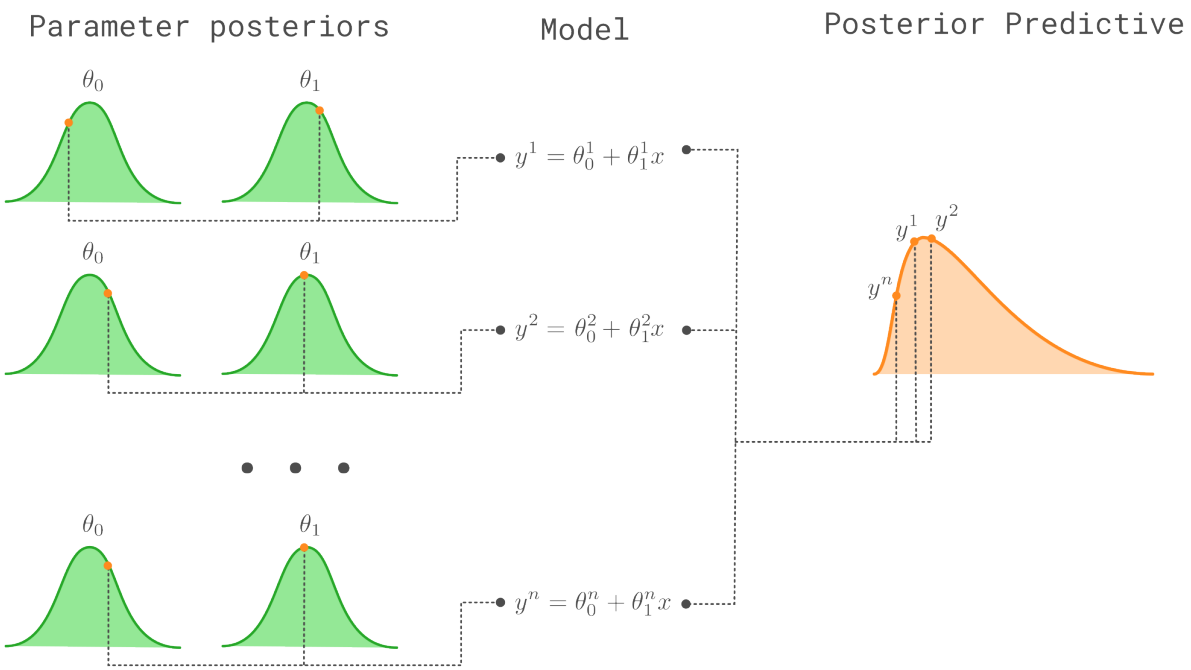

Figure 3.6: Posterior predictive calculation process

Once estimated, the posterior predictive distribution is helpful for several use cases:

[12]Posterior predictive refers to the distribution of the target variable whereas posterior refers to the parameter's distributions.

- In hypothesis testing:
  - Compare the posterior distribution within different groups to infer some characteristics from the population of interest. For instance, to measure the gender gap in salaries, the posterior predictive distribution of salaries between men and women can be compared to check the differences and magnitudes.
- For prediction:
  - Calculate point accuracy metrics for the target, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).
  - Evaluate the model accuracy and performance in both in-sample and out-of-sample scenarios by comparing the posterior predictive distribution with the true distribution.
  - Calculate credible intervals for both the parameters and the target variable to estimate the uncertainty over the target variable through the predictive quantiles (*k-th Percentile of prediction*)
- For simulation:
  - Simulate new data that follows the sample distribution of the target variable (in-sample distribution)
  - Simulate hypothetical scenarios by changing the parameter's distributions and calculate their effect on the target variable, assuming that the same distribution governs the process. For instance, in the case of testing the impacts of a policy change on the gender gap, assessing the effects on salaries.

3.4 Goodness-of-fit and predictive accuracy

Once model parameters are inferred, it is necessary to measure how well the model represents the observed data (in-sample) and its predicting power on new data (out-of-sample).

As reviewed in the posterior section, several point estimate metrics, such as MSE, RMSE, and MAE, can measure the model's predicting power. However, as Bayesian predictions are probability distributions, alternative methods that compare the true and the estimated distributions are preferred, as shown in the following subsections.

3.4.1 The ideal measure: Kullback-Leibler Divergence

Using the theoretical foundations from information theory, the Kullback-Leibler divergence, also known as KL divergence, provides an ideal measure of the discrepancy between two distributions. It measures how much information is lost by using an alternative distribution instead of the true distribution (Lambert, 2018). This measure calculates the "distance" between the true distribution $p(x)$ and the alternative distribution $q(x)$, as shown in Equation (3.8). In this expression, if the distributions $p(x)$ and $q(x)$ are the same, the KL will be 0. Then, the objective is to build a model that minimizes the KL value.

$$KL(p \to q) = \int_{x \in X} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{3.8}$$

$$= \underbrace{\int_{x \in X} p(x) \log(p(x)) dx}_{\text{true distribution}} - \underbrace{\int_{x \in X} p(x) \log(q(x)) dx}_{\text{estimated distribution}} \tag{3.9}$$

The KL divergence can be rewritten as in Equation (3.9), in which the first term is fixed because it is only related to the true distribution, while the second term is related to the model's choice. Given that the first part is fixed, the KL divergence will be minimized when the second term is maximized.

In practice, the calculation of the complete KL divergence can be computationally expensive, given the integral calculations. Hence, the second term in the KL divergence can be used as a proxy to reduce the KL divergence using samples from the posterior predictive distribution. This term, also known as *expected log-pointwise predictive density (elppd)*, is the base for comparing the performance between models and calculating some accuracy measures

such as the WAIC and LOO-CV.

3.4.2 Widely Applicable Information Criterion (WAIC)

One estimate of the *elppd* can be calculated by summing up the log of the average value of the likelihood across the posterior distribution for each data point y_i used to estimate the model (Gelman et al., 2013), as shown in Equation (3.10). Gelman et al. (2013) recommends applying a *bias* correction term that accounts for the uncertainty in the parameter estimation, and it serves as a regularization term that penalizes the model complexity.

$$\widehat{elppd} = \sum_{i=1}^{n} \log\left[E_{\text{posterior}}\left(p(y_i|\theta)\right)\right] - \underbrace{\sum_{i=1}^{n} \text{Var}_{\text{posterior}}\left(\log\left(p(y_i|\theta)\right)\right)}_{\text{bias correction}} \tag{3.10}$$

Then, the WAIC score is calculated as $WAIC = -2\widehat{elppd}$. The lower the WAIC, the better predictive performance of the model18. This score can also be used in model comparison to select the model with the best predictive power. As the WAIC score is calculated with the data points used in the model estimation, it is usually used to measure the model's in-sample predictive power. Nevertheless, if the dataset is split into training, validation, and testing, WAIC calculation using the test set can measure the model's out-of-sample predictive power.

3.4.3 LOO-CV

Using the basic principle discussed in KL divergence, the method of Leave-one-out cross-validation allows the calculation of the model's out-of-sample predictive power by estimating the model with the $n-1$ data points and calculating the $\widehat{elppd}$ for the remaining datapoint to test the model's performance. The LOO-CV score will be the average of all $\widehat{elppd}$. This approach could be computationally expensive for large datasets, as the model will be re-estimated many times.

Vehtari et al. (2015) propose an approximation to the LOO-CV score using samples from the posterior distribution estimated with the entire dataset. Therefore, the model will be

estimated just once. This method is known as *Pareto Smoothed Importance Sampling (PSIS)*.

3.5 Probabilistic programming: A framework to perform MCMC

In practice, Bayesian inference is performed using a programming paradigm called Probabilistic programming. This approach combines traditional programming languages with probability theory to handle uncertainty in the modelling process. The general idea in probabilistic programming is that a model corresponds to a series of interconnected random variables, and all calculations are performed using a graph structure.

There is an ample supply of packages that ease the application of the Bayesian inference framework. Given the author's familiarity with the programming language Python, this thesis uses Numpyro (Phan et al., 2019) as the inference library. One interesting characteristic of Numpyro is that it uses JAX for the sampling process. JAX is a powerful library for efficiently performing numerical computation and tensor operations in CPUs, GPUs, TPUs and other multi-device environments. This characteristic is critical given the computational cost of performing Bayesian inference.

Chapter 4

Data sources

This chapter presents the datasets used in the model estimation and validation. The first section provides details about the data sources and their main characteristics. The second section presents an exploratory data analysis that guides the prior definition and model specification in the next chapter. The final section focuses on the dataset's hierarchical structure, a key component of the model specification.

Several datasets and surveys were evaluated before the model estimation. However, only the Survey of Labour and Income Dynamics (SLID) was considered suitable for developing the model. Table 4.1 presents all datasets reviewed for this work, and the following list explains the main reasons for using the SLID survey over the existing datasets.

- Cross-sectional labour market surveys such as the LFS are designed to provide a most representative view at the population level and to track general trends in the labour market. However, the longitudinal nature of SLID provides a more detailed view of the individuals and their dynamics within the labour market, such as transitions, durations, and occurrences of individual financials and work situations (StatisticsCanada, 2012).
- SLID provides more detailed information about the characteristics of the individuals, such as age, prior experience, education level, and other characteristics associated with the job. In contrast, LFS age and education level are grouped into fewer categories.

LFS does not report previous work experience, which, according to the literature, is one of the most important predictors of a worker's salary.

- Salary in SLID is reported annually, which provides a better estimate of the annual income from labour sources. Conversely, salaries in the LFS are reported as hourly wages, which can be difficult to calculate for many full-time workers.
- Given that the LFS panel rotates every six months, and a panel is added every month, the salary estimates, and other representative variables can vary between periods. In contrast, the SLID panel is more stable and can produce more reliable estimates for the same variables.
- Although the SLID dataset could be considered outdated when this thesis is written, labour market changes take longer. Therefore, this dataset still contains essential information about the dynamics of the labour market that can be applied today. One argument supporting its use is that the Industry and Occupation classification system has had only minor changes and additions at the subcategory level. However, the structure remains the same as the one used in the SLID survey[1].

[1] In 2022, Statistics Canada released a major change in the occupation classification system that organizes occupations in TIERS according to the education and skill level required for the job. However, these TIERS are just a subgroup of the original NOC categories used in this thesis.

Dataset	Description	Frequency	Reference periods
Labour Force Survey (LFS)	A cross-sectional survey that estimates the state of the labour market in Canada	Monthly	1987-2022
Survey of Labour and Income Dynamics (SLID)	A longitudinal survey focused on understanding the labour market activity, household income, and the changes experienced by individuals and families through time	Annual	1994-2011 (Discontinued)
Survey of Older Workers (SOW)	A cross-sectional survey to identify the factors that influence the decision to retire or remain working.	One time	2008
Canadian Income Survey (CIS)	A cross-sectional survey to monitor the income and sources of individuals and their household characteristics	Annual	2012-2018
Employment and Insurance Coverage Survey (EICS)	Cross-sectional survey monitoring the Employment Insurance (EI) program. This survey provides an overview of the socio-economic characteristics of the unemployed and those not in the labour force. Also, it covers maternity and parental benefits.	Annual	2007-2018

Table 4.1: Review of publicly available data sources related to the Labour Market

4.1 Survey of Labour and Income Dynamics - SLID

The Survey of Labour and Income Dynamics (SLID) was a statistical program developed by Statistics Canada between 1993 and 2011 to investigate the Canadian population's income sources. This survey provides an additional perspective to the labour market surveys focused on the individual's characteristics. After 2011, this survey was discontinued, and some modules were included in the monthly Labour Force Survey.

The SLID is a household longitudinal survey applied across all provinces and territories in Canada and collected by telephone. The sample, a subsample of the Labour Force Survey, consists of two panels of households of around 17,000 households and 34,000 respondents interviewed once a year for six years. A new panel was introduced every three years, so there was always an overlap of three years (StatisticsCanada, 2012).

Although the original SLID is a longitudinal survey, the Public-Use Microdata File (PUMF) available on the Statistics Canada website corresponds to the cross-sectional version of the survey. That means all identifiers that track individuals within the panel are not provided. Statistics Canada applies this transformation to preserve data privacy and confidentiality. Similarly, geographic information is aggregated at the province level. Given that the wage model proposed in this thesis is part of the ILUTE model, the SLID dataset is filtered to include samples from Ontario to match the geographical context of the ILUTE model.

SLID dataset consists of a set of 147 variables in its last version (2011). These variables are grouped into the following five categories that represent some characteristics of the individual, their household, or their labour status at the time of the survey. According to the theoretical foundations discussed in the literature review, these variables are filtered to select the most relevant for modelling salaries. Table 5.1 details the variables used for the model specification.

Variable group	Variable subgroups	Variables used in the model
Sample	Identifiers, Sample variables	Personal identifier, Survey year, Sample weight.
Personal	Demographics, Ethnocultural characteristics, Activity limitations, Geography, Family and Household characteristics	Age, Sex, Major activity in reference year, Province of residence
Labour	Labour market activity patterns, Work experience, Job characteristics	Labour Force Status, Work experience, Job duration, Wages and benefits, Occupation, Industry, Job tenure, Self-employment, Employment sector (public or private)
Financial situation	Income sources	Primary income source, Annual Salary
Education	Educational activity, Level of schooling	Highest level of education, Student status

Table 4.2: Variables used in the model specification

4.2 Hierarchical structure: Industry and Occupation

Labour markets are inherently hierarchical. Workers and firms are organized and classified into different industries that group the purpose and common operations of a business within an economy. Similarly, workers can be classified into occupations based on the tasks and skills required to perform the job. In terms of data structure, this relationship can be defined as one-to-many. One industry can contain many occupations that cover different processes and tasks. This hierarchy structure defines the category each firm or worker belongs to and

other aspects of the labour market. Wage differentials are highly correlated with the industry and occupation of the worker.

Statistics Canada, jointly with the USA and Mexico statistical departments, developed the North American Industry Classification System (NAICS) and the National Occupation Classification (NOC), which are the systems to classify industries and occupations in the North American context[2]. These systems have evolved continuously to the last version released in 2021 for NAICS and NOC[3]. Although the SLID survey used the NAIC and NOC 2001 versions, the dataset used in this thesis was translated into the last versions to ensure future validity and updates[4].

The NAICS system classifies industries into ten principal categories that can be subdivided into multiple subcategories. Similarly, the NOC system classifies occupations into ten large groups divided into multiple categories. These classification systems facilitate the standardization and aggregation of similar industries or occupations. Given the dynamic nature of the labour markets, these systems can introduce any changes in the short, medium or long term. However, defining the aggregation level requires a balance between the representation detail (heterogeneity), model performance, and avoiding aggregation biases.

Hence, this thesis adopts an aggregation of 16 industries (Table 4.3) and 25 occupations (Table 4.4) commonly used in several labour market reports from Statistics Canada. Although the original implementation in ILUTE only uses eight industries and occupations, this proposal can be translated into the existing aggregation in ILUTE and into the NAICS and NOC2021 versions.

[2]NOC is a Canadian adaptation of the SOC system applied in the USA, therefore, it is only used in Canada.

[3]Most of the changes between these versions are in the subcategory level, which reflects the inclusion of new businesses and occupations into the existent industry and occupation codes at the principal category level.

[4]SLID reports industry and occupation codes at the most upper level (level 1), therefore, the changes in the last version of the system does not affect them.

Code	Industry	Code	Industry
0	Agriculture	**8**	Finance and Real State
1	Forestry, Oil, Mining	**9**	Professional, Scientific and Technical
2	Utilities	**10**	Business support
3	Construction	**11**	Educational services
4	Manufacturing	**12**	Health care and social services
5	Trades	**13**	Accommodation and Food services
6	Transportation and Warehousing	**14**	Other services
7	Information and Culture	**15**	Public Administration

Table 4.3: Industry categories in the proposed model

Code	Occupation	Code	Occupation	Code	Occupation
0	Senior manager	**8**	Teacher/Professor	**16**	Clerks/Cashier
1	Middle management	**9**	Government/Social services	**17**	Construction trades
2	Business/Finance Professional	**10**	Protective services	**18**	Transport/Equipment operator
3	Secretarial/ Administrative	**11**	Child care/Home support	**19**	Trade contractor/supervisor
4	Natural/Science professional	**12**	Art/Culture occupation	**20**	Trade helper/labourer
5	Technical specialist	**13**	Clerical/Supervisor	**21**	Other trades
6	Health professional	**14**	Chef/Food services	**22**	Operator/Assembler
7	Health Assistant	**15**	Sales/Service	**23**	Manufacturing labourer

Table 4.4: Occupation categories in the proposed model

4.3 Exploratory Data Analysis

This section briefly explores the filtered dataset (variables listed in Table 5.1). All analyses are focused on understanding the data generation process and providing clues for guiding the model construction in the following chapters.

Given that SLID covers multiple years, all monetary values must be converted to real terms by eliminating the effect of changes in consumer prices (inflation or deflation) to compare monetary values at different times correctly. This task is performed by converting salaries into real terms (2023 Canadian dollars) using the Consumer Price Index dataset from Statistics Canada. As shown in Figure 4.1, salaries are positive and right-skewed distributed with a mean of $67.100 and median around $60.300. As expected, most salaries are concentrated around the mean, and the density decreases as salary increases. This shape resembles the shape of continuous distributions such as Gamma or log-normal, which are common choices used by the insurance and financial sector to model income and salaries (Nielsen, 2010).

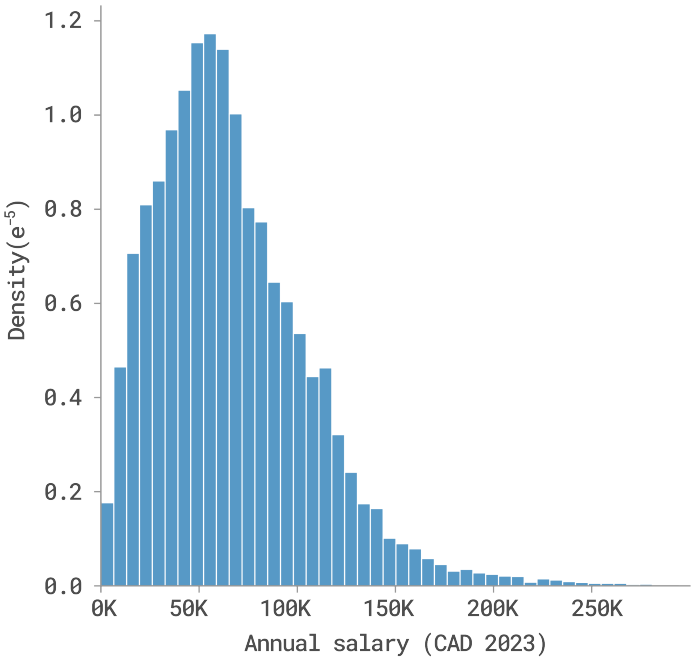

Figure 4.1: Salary distribution for the SLID dataset

The comparison between average salaries by industry and the average salary in Ontario is shown in Figure 4.2, in which both did not have significant changes during the SLID period. Some exceptions to this behaviour are the Public Administration and Utilities industries, which had an increase of 32% and 30% in real terms, respectively.

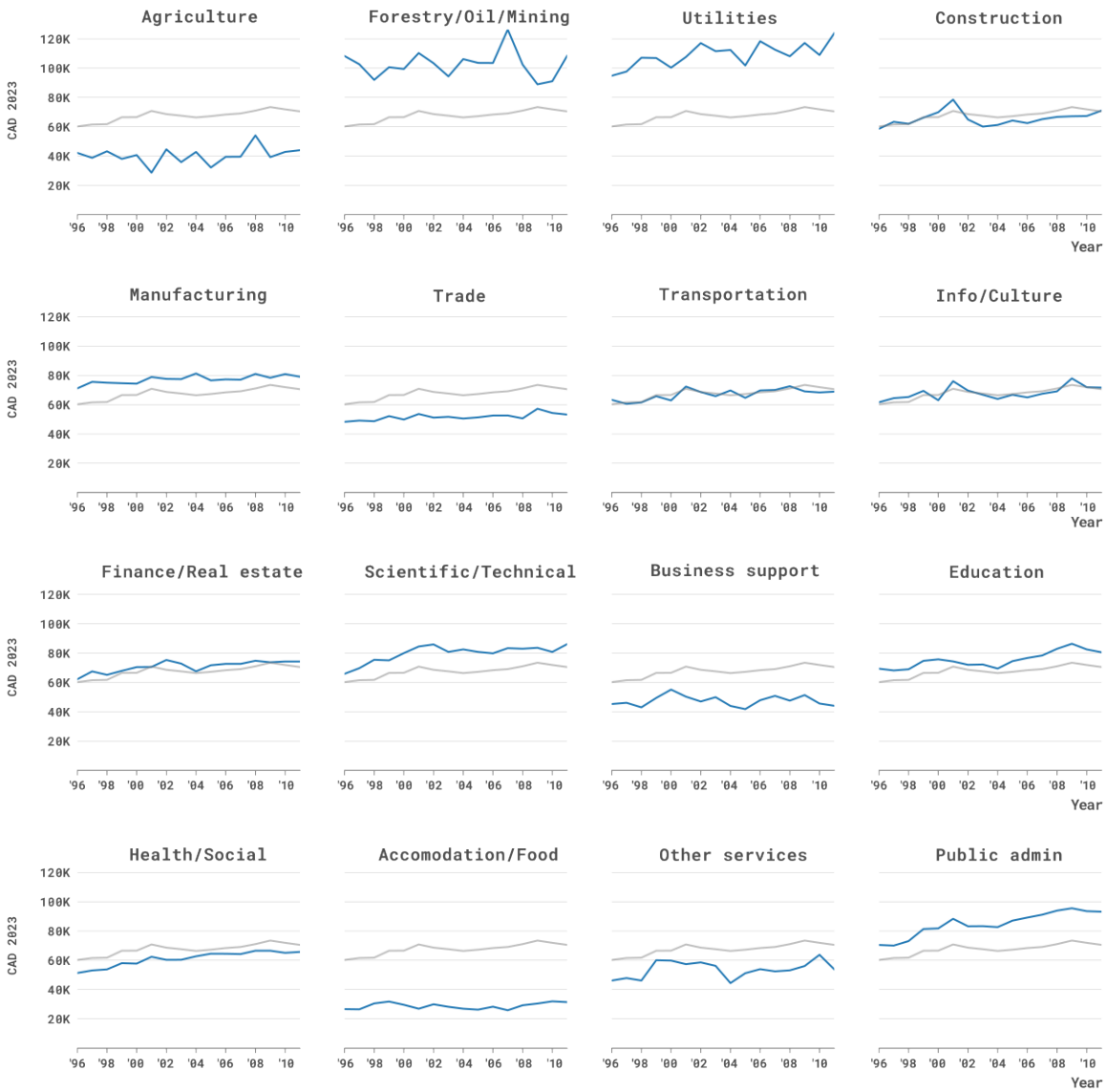

Figure 4.2: Average salary by industry in real terms (2023 Canadian dollars)

The fact that salaries in real terms did not significantly change allows the construction of models that do not consider the autoregressive nature of time series, which simplifies the modelling process and improves the interpretability of the model.

According to the theoretical model in the literature review chapter, education level is one of the most important predictors of a person's salary because it is a proxy measure of the skill set necessary to perform a job. Figure 4.3 shows how salary increases as the education level increases.

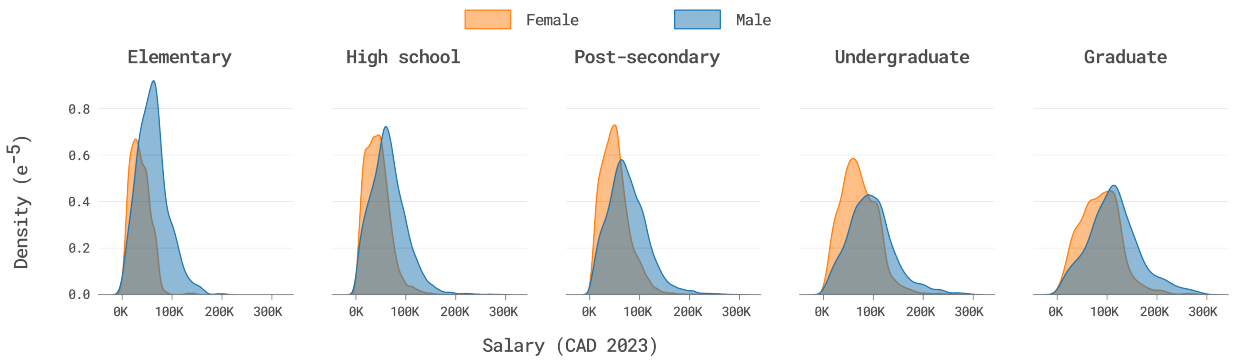

Figure 4.3: Salary Distribution by education level and gender

Although education level can explain salary increases, other variables such as age, years of experience, and tenure time at the current or previous jobs have also been associated with the salary level. As some industries and occupations require specific skills learned on the job or through many years of experience, the experience level, age or job tenure could provide more information than only using the education level. Figure 4.4 presents the pairwise relationship between these variables (first rows) and the salary (last row).

As expected, age, experience, and tenure seem to be positively correlated. However, Figure 4.4 shows a weak linear relationship between the variables of interest and salaries, with a high variance across all data points. This variability can be explained by the hierarchical structure of the data, in which salaries are also determined by the industry and occupation.

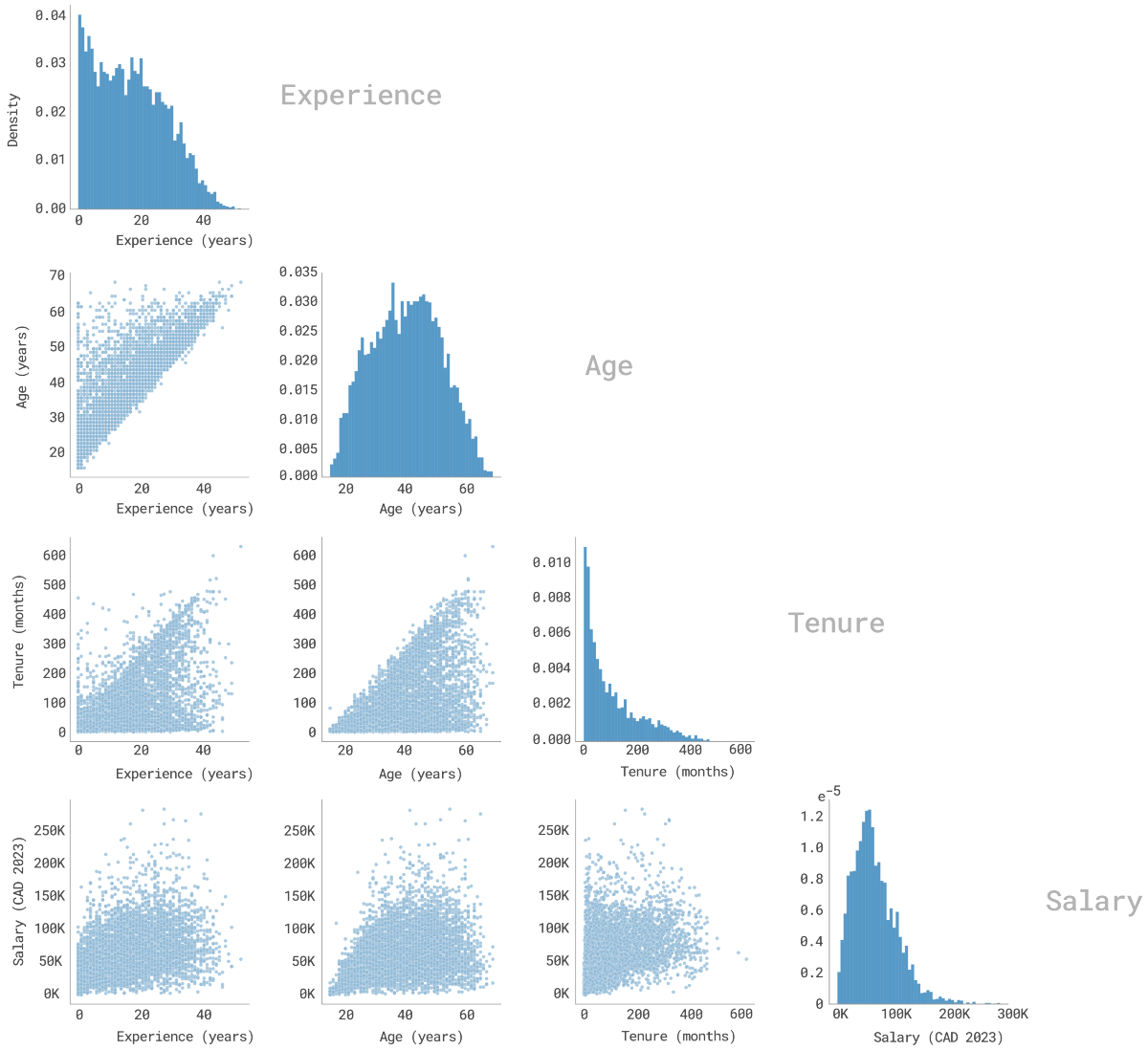

Figure 4.4: Salaries by experience level, age, and tenure

When data is filtered by industry and occupation, the linear relationship between experience, age, tenure, and salary becomes more explicit, as shown in Figure 4.5. The following chapter will analyze this hierarchy in more depth to propose a model that captures this structure in the data.

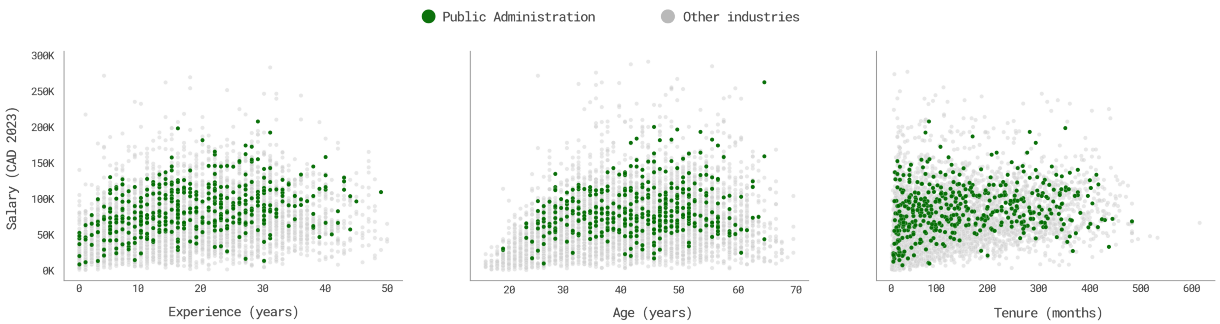

Figure 4.5: Salary distribution by several attributes in the Public Administration industry

In addition to the variables supported by the theoretical models, some individual characteristics such as union participation, job sector (public or private), and employment type (self-employee or employee) were compared to assess their effect on salaries, as shown in Figure 4.6. Previous studies have found evidence that the effect of bargaining increases salaries because unions have more negotiating power (Lewis, 1986; Mcdonald and Solow, 1992). This negotiating power increases in some monopolistic sectors, such as the public sector, because of the lack of competition (Gunderson, 1979). In contrast, self-employees lack employee benefits and stability, which is directly related to lower salaries (Hamilton, 2000).

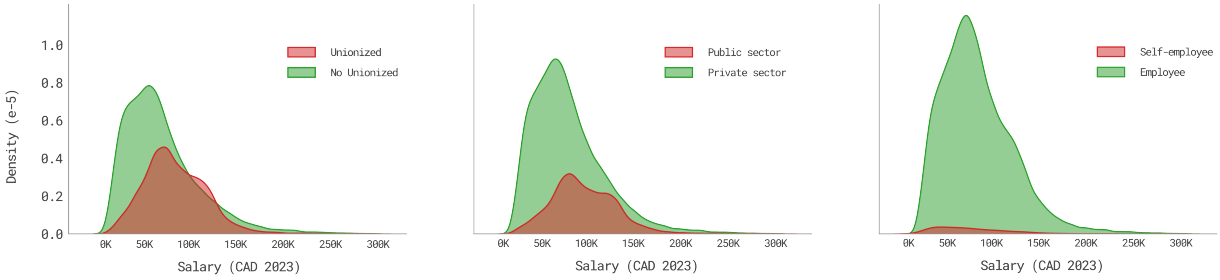

Figure 4.6: Salary distribution comparison between union, sector, and employment type.

Chapter 5

Salary model estimation

This chapter details the construction of a salary model that predicts annual salaries based on the individual attributes of a worker using the Bayesian inference framework. This model is an update of the current implementation of the wage model in the labour market module of the ILUTE framework presented in Section 2.2.1.

The content in this chapter is structured into seven subsections: the first part discusses the importance of the hierarchical structure in the model definition. Then, the second section presents the variables used in the model specification. The third section details the model specification based on the findings in the exploratory data analysis in Section 4.3. The fourth section presents a set of proposed models and the evaluation framework to select the optimal model. The fifth section discusses the model interpretability. The sixth section analyzes the stability of parameters over time, and the final section presents the model validation on data that was not used in the model estimation (out-of-sample).

5.1 Data structure: from single to multilevel structure

The previous sections discussed how a linear combination of some personal attributes can explain salaries. Additionally, it explored how salary variability is attributable to the hierarchical structure defined by the industry and occupation categories. Therefore, exploring the

data structure and data generation process before modelling salaries is important for defining the best approach.

Considering these factors, a simple linear combination of personal characteristics, also known as a ***Pooled model***, can be one way to model salaries. In this approach, all industries and occupations are equally treated, which implies that salaries are modelled as the average for all industries and occupations. Although theoretically correct, this model type fails to represent the structure of the industry-occupation relationship and produces poor goodness-of-fit and low performance in the long run. One example of this approach is a linear regression that predicts salaries using the data from the whole labour market. This model will predict the average salary but certainly will overestimate the value for low-skill jobs and underestimate salaries for high-skill jobs.

One way to improve the pooled model performance is to estimate one linear model for each hierarchy level. That means estimating one linear regression for each industry-occupation combination in the labour market. This approach, also known as the ***No-Pooled model***, captures the hierarchical structure with the downside that it treats each industry-occupation combination as totally independent from the other ones. This independence assumes that salaries in the same industry but in different occupations are unrelated, contrary to the nested characteristic observed in the exploratory data analysis section.

This independence results in a model that is less robust to outliers, prone to overfit the data, and hard to meet the statistical significance for industry-occupation categories with a small number of data points. Ultimately, these disadvantages affect the model's performance and the goodness-of-fit in the no-pooled approach. Therefore, an implementation that accounts for the hierarchy in the data and the internal relationships is needed to predict and model salaries adequately.

Taking the no-pooled approach as a base, this model can be improved by introducing some dependency between industry-occupation classes. The ***Multilevel or Hierarchical model*** is an approach between the last two model types. It represents each industry-occupation combination as a separate model but draws information from higher levels (industry level) to produce more realistic predictions. By using information from higher levels, the model is

more robust to outliers, less prone to overfit, and produces more realistic predictions because it considers the inner relationships.

One interesting way to understand the advantages of using the hierarchical model over the other approaches is through the *Simpson's paradox*. Using the synthetic salary and experience dataset shown in Figure 5.1 with some artificial categories (one industry and three occupations within that industry), the pooled model (grey line) predicts that salaries decrease with experience. In contrast, the no-pooled model (dashed lines) captures the expected behaviour in which salary increases with experience. The Hierarchical model (solid lines) provides the most reliable estimation in this conceptual demonstration because it captures the expected behaviour and is less sensitive to outliers. This performance improvement comes with an increase in computational costs.

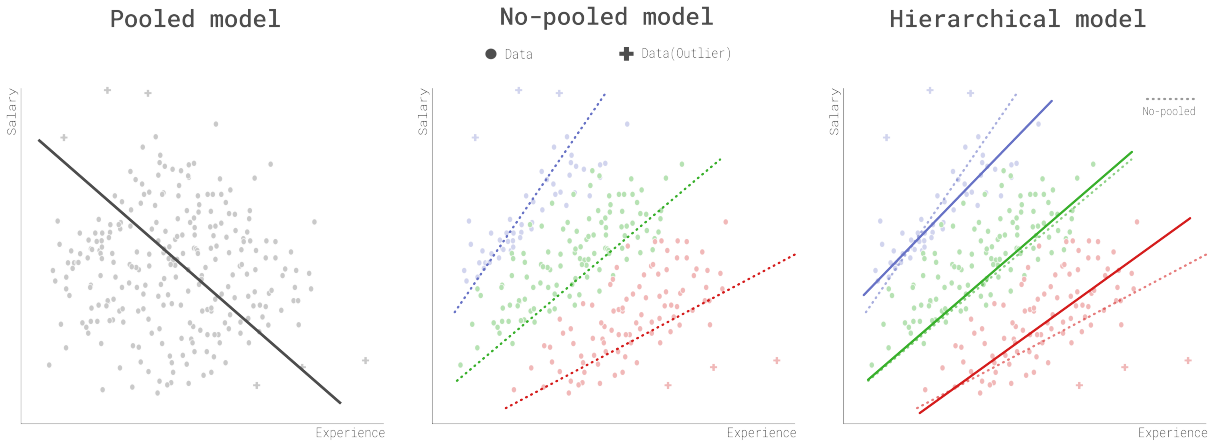

Figure 5.1: Effect of model structure on the data representation

Given that the hierarchical approach draws information from the higher level (industry level), salaries predicted for each category tend to move towards the average industry value, represented by the dashed lines in Figure 5.1. This behaviour is known as *Shrinkage*, which can be interpreted as a regularization process that increases the model's robustness by reducing the effect of outliers and the risk of overfitting.

As a brief overview, the following list and Figure 5.2 summarize the differences in the inference process for the three approaches[1]

- **Pooled model**: In this approach, the modeller defines a set of priors for each model parameter based on the existing knowledge. Then, the posterior distribution for each model parameter is estimated using the Bayesian framework reviewed in Chapter 3. The posterior distributions represent the average value of each parameter for all industries and occupations.

- **No-pooled model**: In this approach, each industry-occupation combination has a set of priors for each model parameter. Each prior is defined by the modeller based on the existent knowledge. Then, the estimation result is a distribution matrix for each model parameter. Each distribution in this matrix corresponds to the estimated posterior distribution for that specific industry occupation and that given parameter.

- **Multilevel or Hierarchical model**: This approach follows the same process as the No-pooled model, adding an extra step before setting the priors. One of the key characteristics of the hierarchical models is the transfer of information between higher and lower levels. This is achieved by setting a set of hyperpriors that defines the localization and scale of the prior distributions (prior parameters). In the hierarchical approach, the modeller only defines the shape of the prior distribution and lets the data select the best parameters for that prior based on the hyperpriors. After the estimation process, a matrix of distributions for each model parameter is generated. Each one of these distributions corresponds to the estimated posterior distribution for that specific industry occupation and for that given parameter. However, each distribution shares information with other distributors in the same industry. The estimation process also provides posterior distributions for the hyperpriors, which corresponds to the distribution of the information shared across categories in the same industry. This process makes the hierarchical approach robust and powerful for modelling complex data structures.

[1]There are similar model structures in the frequentist approach. The pooled and no-pooled counterpart could be a set of simple linear regressions or ANOVA models, and the Hierarchical model is known in the frequentist framework as a mixed-effects model. Numerous examples also exist in the random utility / discrete choice modelling literature.

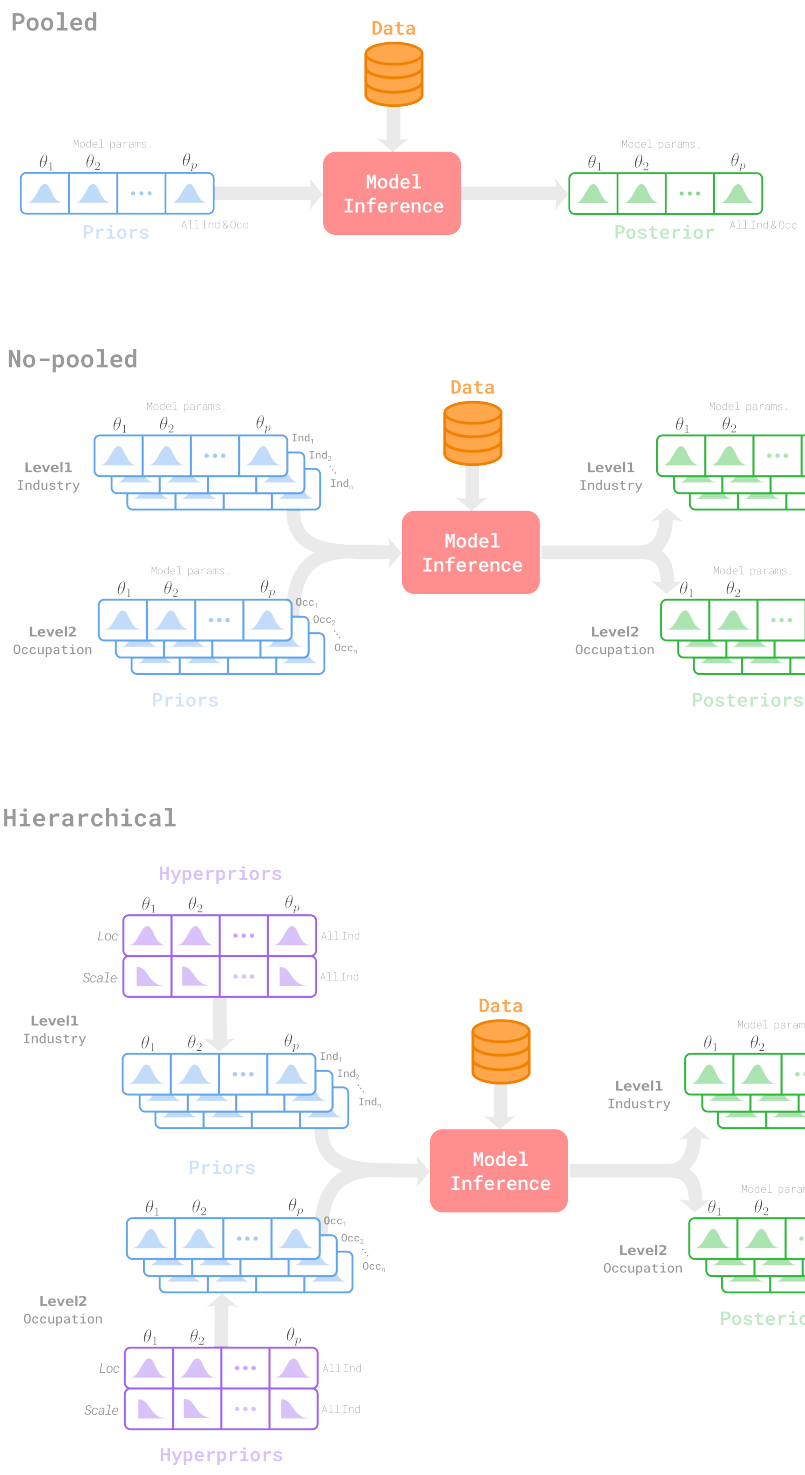

Figure 5.2: Inference process by model structure

Despite these approaches exemplified in this section using linear models, it is important to highlight that these structures are model agnostic, which means that they can be used with linear and non-linear models. Therefore, the three approaches reviewed in this section only define the model structure.

Based on these arguments, the Hierarchical structure seems to be the best option for modelling salaries using the SLID dataset. However, the model specification and selection sections compare the three approaches to measure their performance using the SLID data. In the model selection, the optimal model is chosen based on the measures reviewed in Section 3.4.

5.2 Model variables

The variables used in the model estimation are detailed in Table 5.1. It is important to highlight that all variables are used to select the model structure. After choosing the optimal structure, the final variables are selected through a Forward selection process, as presented in Section 5.4.

Based on the comments of Harmon (2013), the underestimation of the existent wage model could be related to the quality of the hourly salary data. Therefore, the proposed model uses annual salary as the target variable instead of the hourly wage.

The original SLID dataset is split into two: from 1996 to 2007 and 2008 to 2011. The first corresponds to the dataset used in the model estimation, and the last is the dataset used in the validation section.

Name	Units	Type	Description
Experience	years	Continuous	Years of previous experience in the occupation and industry
Sex	0 or 1	Categorical	Dummy variable that takes 0 for Males or 1 for Females
Education Level	0 or 1 for each education level: No education, Elementary, High-school, Postsecondary, Undergraduate, Graduate	Categorical	The dummy variable represents the education level using a One-hot encoding.
Age	years (15-99)	Continuous	Age of the individual at the time of the survey
Tenure	months	Continuous	Number of months in the same job
Union	0 or 1	Categorical	Dummy variable that takes 1 if the individual is unionized or 0 otherwise
Public sector	0 or 1	Categorical	The dummy variable takes 1 if the individual works in the public sector and 0 otherwise.
Self-employment	0 or 1	Categorical	The dummy variable takes 1 if the individual is self-employed and 0 otherwise.

Table 5.1: Variables used in the model specification

5.3 Model specification

Once the model structure and data variables are defined, it is necessary to specify the relationships between the variables of interest in the model by recreating the data generation process. Two of the most relevant characteristics of the data generation process are the distribution of the target variable and the relationships between the explanatory variables.

As discussed in the exploratory data analysis section, salaries are positive and right-skewed distributed, which resembles the Gamma distribution. On the other hand, theoretical models explored in the literature review section show how the salary of a particular worker can be explained by using a linear combination of different personal attributes such as education level, experience, gender, and age, among others. Hence, the model specification must meet these target distribution and linearity requirements.

A model that meets these requirements is the *Gamma Generalized Linear Model (Gamma GLM)*, which belongs to the family of linear models that represents a process in terms of a linear combination with error distribution different from the normal distribution, as in the ordinary linear regression (Nielsen, 2010).

Given a set of explanatory variables $X = [X_1, X_2, ..., X_p]$ and a set of model parameters $\theta = [\theta_0, \theta_1, ..., \theta_p]$, the Gamma GLM is defined by the following components:

- **Random component**: The response variable Y follows a Gamma distribution (Equation (5.1)) defined by the parameters α and β (shape and scale respectively). The expected value of this distribution is $\mu = \frac{\alpha}{\beta}$

$$f(y, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, y > 0 \tag{5.1}$$

- **Systematic component**: The linear combination η of explanatory variables X and the model parameters θ.

$$\eta = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_p X_p \tag{5.2}$$

- **Link function**: This function connects the expected value of the target variable with the linear predictor η using a log link and ensures that the linear predictor is positive and continuous. When solved for μ, it provides the expectation over the target variable.

$$\begin{aligned} log(\mu) &= \eta \\ \mu &= e^{\eta} = e^{\theta_0+\theta_1 X_1+...+\theta_p X_p} \end{aligned} \tag{5.3}$$

Hence, the inference process is focused on estimating the parameters θ's in the systematic component, whereas the parameters α and β in the random component are inferred indirectly using the systematic component and the log-link function. The following subsections present the model specification of the Gamma GLM for each model structure (Pooled, No-Pooled, and Hierarchical).

5.3.1 Pooled model

Equation (5.4) presents the model specification for the Pooled structure and Figure 5.3. Model graph - Pooled shows the model graph. The expression Equation (5.4)-[1] corresponds to the random component, while the expressions Equation (5.4)-[2] and Equation (5.4)-[4] correspond to the priors for the model parameters θ_p and α. The model parameters correspond to all variables listed in Table 5.1.

$$\begin{aligned} Y &\sim Gamma(\alpha, \beta) && [1] \\ \alpha &\sim Uniform(0, 100) && [2] \\ \beta &= \alpha/\mu && [3] \\ \theta_p &\sim Normal(0, 1) && [4] \\ \mu &= e^{\theta_0+\theta_1 X_1+...+\theta_p X_p} && [5] \end{aligned} \tag{5.4}$$

The choice of the priors is given by:

- For the parameters θ's, the prior selected is the normal distribution with mean 0 and

sigma 1. The mean centred in 0 ensures that the parameters can take positive or negative values according to the relationships in the data.

- For the parameter α, the prior selected is the uniform distribution with lower limit 0 and upper limit 100. The uniform distribution is a non-informative prior that ensures different values for α are tested with equal probability of occurrence. This choice is made because there is no evidence of the α value in previous salary prediction exercises.
- The parameter β does not require a prior because it is calculated based on the other parameters.

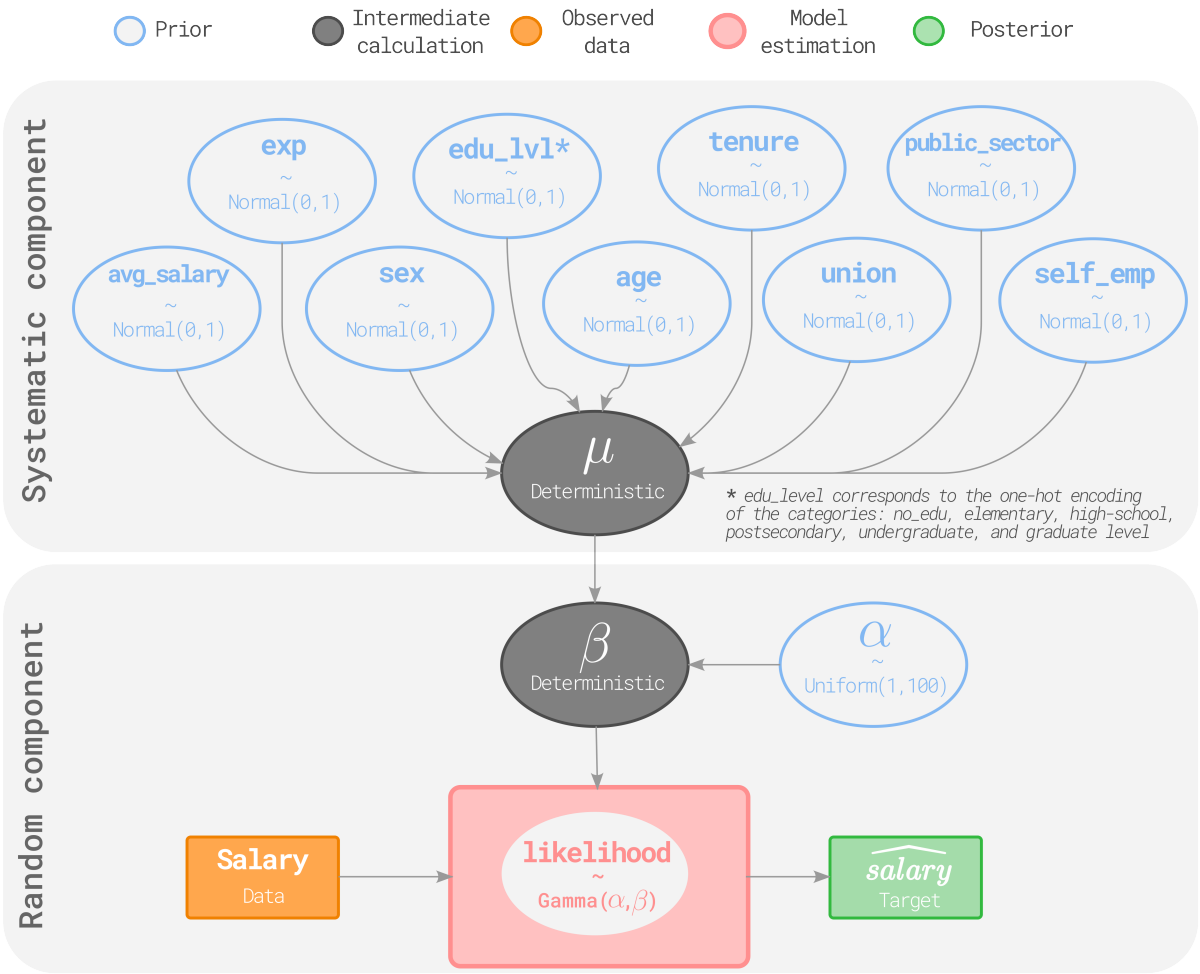

Figure 5.3: Model graph - Pooled

5.3.2 No-pooled model

Equation (5.5) presents the model specification for the No-pooled structure, and Figure 5.4 shows the model graph. The model parameters correspond to all variables listed in Table 5.1. The superscripts *ind* and *occ* correspond to each hierarchy level (industries and occupations) presented in Section 4.2.

$$
\begin{aligned}
Y &\sim Gamma(\alpha, \beta) && [1] \\
\alpha &\sim Uniform(0, 100) && [2] \\
\beta &= \alpha/\mu && [3] \\
\theta_p^{ind} &\sim Normal(0, 1) && [4] \\
\theta_p^{occ} &\sim Normal(0, 1) && [5] \\
\eta^{ind} &= \theta_0^{ind} + \theta_1^{ind} X_1 + ... + \theta_p^{ind} X_p && [6] \\
\eta^{occ} &= \theta_0^{occ} + \theta_1^{occ} X_1 + ... + \theta_p^{occ} X_p && [7] \\
\mu &= e^{\eta^{ind} + \eta^{occ}} && [8]
\end{aligned}
\tag{5.5}
$$

Like in the Pooled model, the choice of the priors in the No-pooled model is given by:

- For the parameters θ's for both *ind* and *occ*, the prior selected is the normal distribution with mean 0 and sigma 1.
- For the parameter α, the prior selected is the uniform distribution with lower limit 0 and upper limit 100.
- The parameter β does not require a prior because it is calculated based on the other parameters.

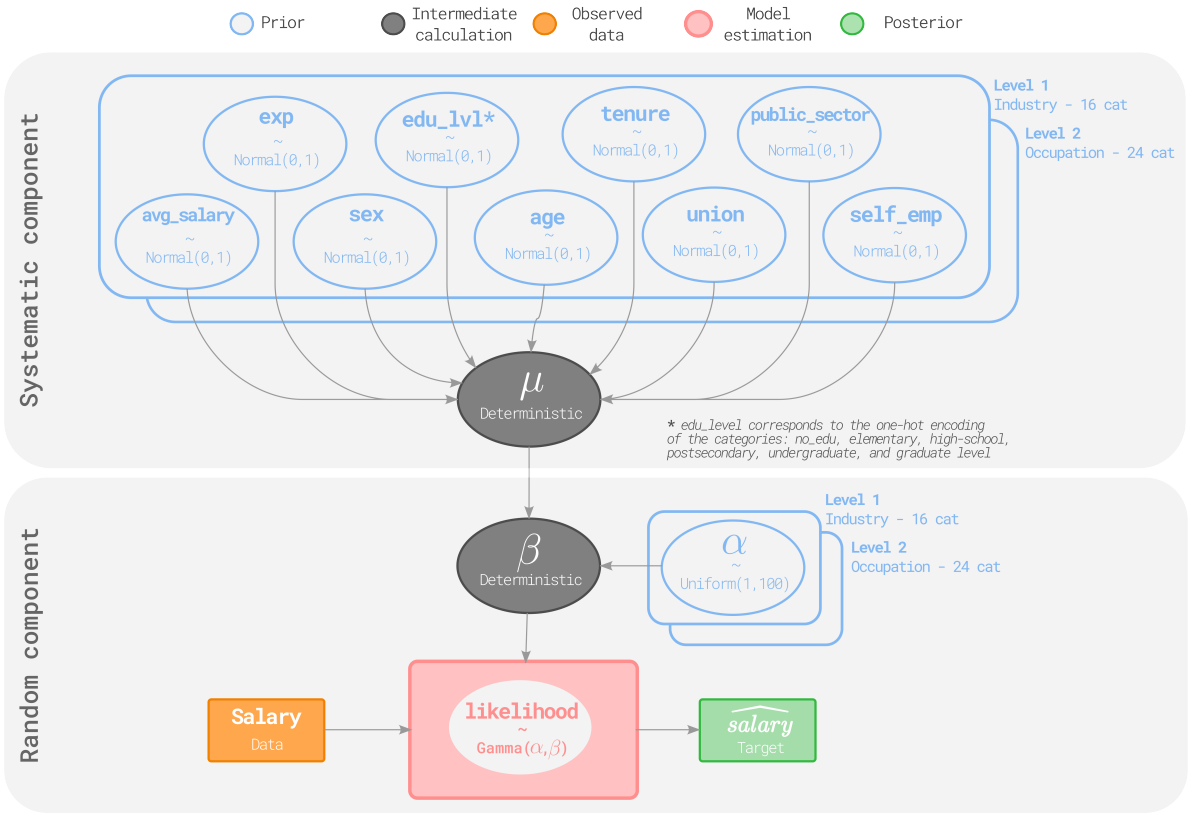

Figure 5.4: Model graph - No-pooled

5.3.3 Hierarchical model

Equation (5.6) presents the model specification for the Hierarchical model, and Figure 5.5 shows the model graph. The hierarchical model uses the variables listed in Table 5.1 and superscripts *ind* and *occ* presented in Section 4.2. However, it adds two hyperpriors (one for the localization parameter and one for the scale parameter) that share information from a higher to a lower level. These hyperpriors are defined for both the industry and occupation level.

$$
\begin{aligned}
Y &\sim Gamma(\alpha, \beta) && [1] \\
\alpha &\sim Uniform(0, 100) && [2] \\
\beta &= \alpha/\mu && [3] \\
loc_p^{ind} &\sim Normal(0, 1) && [4] \\
scale_p^{ind} &\sim HalfNormal(1) && [5] \\
loc_p^{occ} &\sim Normal(0, 1) && [6] \\
scale_p^{occ} &\sim HalfNormal(1) && [7] \\
\theta_p^{ind} &\sim Normal(loc_p^{ind}, scale_p^{ind}) && [8] \\
\theta_p^{occ} &\sim Normal(loc_p^{occ}, scale_p^{occ}) && [9] \\
\eta^{ind} &= \theta_0^{ind} + \theta_1^{ind} X_1 + ... + \theta_p^{ind} X_p && [10] \\
\eta^{occ} &= \theta_0^{occ} + \theta_1^{occ} X_1 + ... + \theta_p^{occ} X_p && [11] \\
\mu &= e^{\eta^{ind} + \eta^{occ}} && [12]
\end{aligned}
\tag{5.6}
$$

The choice of the priors in the Hierarchical model is given by:

- For the hyperprior loc_p^*, the distribution selected is the Normal distribution with mean 0 and sigma 1. That allows the model to choose the best value according to the observed data and share the same distribution for occupations in the same industry.
- Likewise, the hyperprior $scale_p^*$ is defined with the Half-Normal distribution with sigma 1. This distribution ensures that all values are positive, which is mandatory given that $scale_p^*$ is the sigma value in the correspondent prior θ_p^*.
- For the parameters θ's for both *ind* and *occ*, the prior selected is the normal distribution with mean loc_p^* and sigma $scale_p^*$.
- For the parameter α, the prior selected is the uniform distribution with lower limit 0 and upper limit 100.
- The parameter β does not require a prior because it is calculated value.

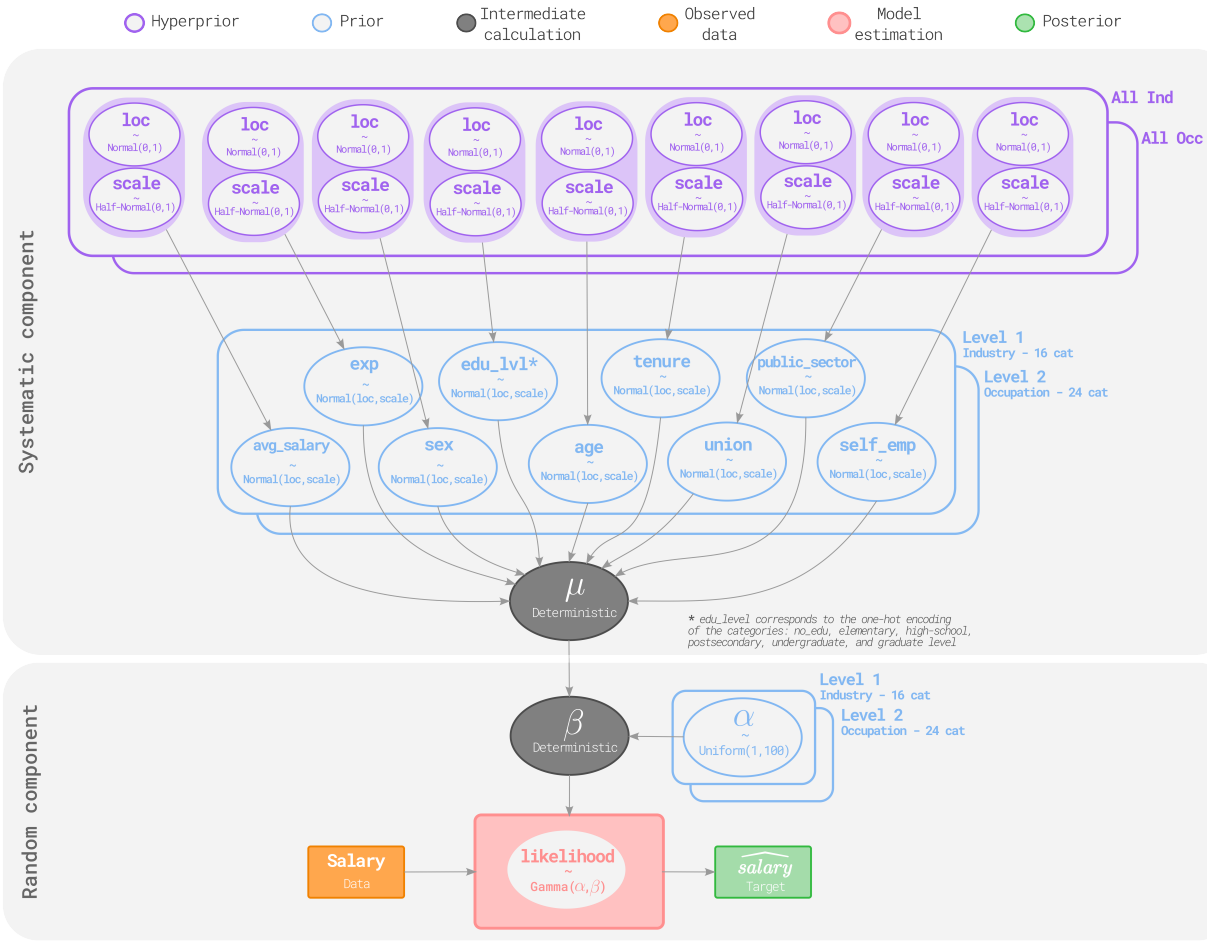

Figure 5.5: Model graph - Hierarchical

5.4 Model selection

After running the inference on the three model structures, the results are compared using the performance metrics presented in Section 3.4. The model selection is divided into two sequential parts: 1) the model structure selection, which selects the best performing model between the pooled, no-pooled, and hierarchical structure, and 2) based on the structure selection, this part focuses on the most relevant variables based on model performance. The result from the 2-stage selection process is the optimal model, which is the proposal to replace the existing wage model in the labour market module of the ILUTE framework.

5.4.1 Structure selection: Pooled vs. No-pooled vs. Hierarchical model

Figure 5.6 shows a model comparison using the LOO-CV measure (ELPPD score) and the WAIC score. The maximization of the ELPPD minimizes the KL divergence. Hence, in this figure, the higher the elppd value, the better the model.

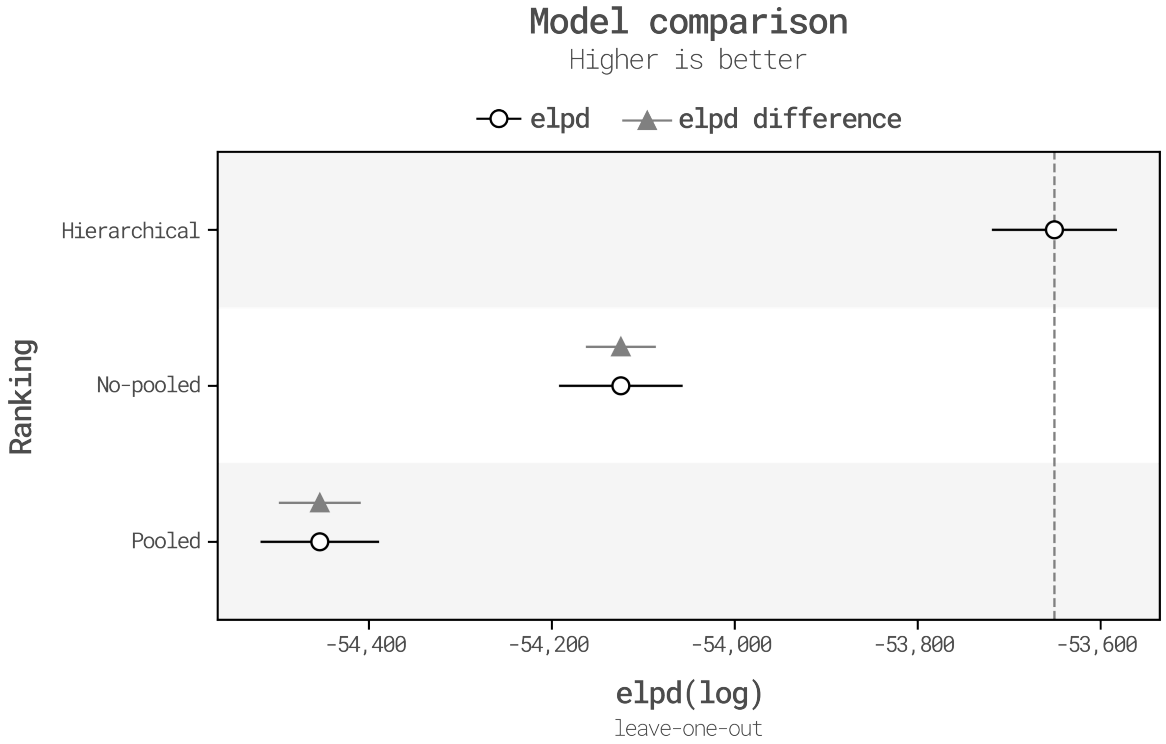

Figure 5.6: Model structure comparison

As expected, the hierarchical model shows better performance than the other approaches. The white circle represents the point estimate of the ELPPD-LOO score that measures the out-of-sample model performance. In contrast, the horizontal lines represent the uncertainty or error bars for the point estimate. Likewise, the ELPD difference provides a measure of comparison between each model and the reference model (best performant). The in-sample point estimate measures the model performance with the data used in the model estimation.

The in-sample ELPPD score is an estimate of the performance of the training dataset. In contrast, the ELPPD-LOO score is an estimate of the performance on the validation dataset, which is a better estimate of the model performance in a real setting. The comparison between these two values ensures that the model is not overfitting. As expected, the model's performance in the in-sample case (training data) is higher than the out-of-sample case (validation data), which implies that the *shrinkage effect* of the hierarchical approach is improving the performance by reducing the model complexity and avoiding overfitting.

The results in Figure 5.6 confirm that the hierarchical structure provides a higher predictive power than the pooled and no-pooled approach. Including the hyperpriors allows the model to better capture the complex relationships and dependencies in the data compared to the other approaches (pooled and no-pooled). Therefore, from this point, all tests and additional runs are performed using the hierarchical structure.

5.4.2 Forward variable selection

After selecting the optimal model structure, a Forward selection process is performed on the variables set to choose the most relevant from the predictive perspective. This process starts from the most straightforward model representation to the total variable set by adding one variable at each step. The simplest model corresponds to an intercept model (fixed effects), representing the average salary for that industry and that occupation.

Figure 5.7 compares each variable addition from the simplest model to the one with the complete set of variables. These models are ranked from the best to the worst in terms of model performance, using the same ELPPD-LOO and WAIC scores used in the model structure selection.

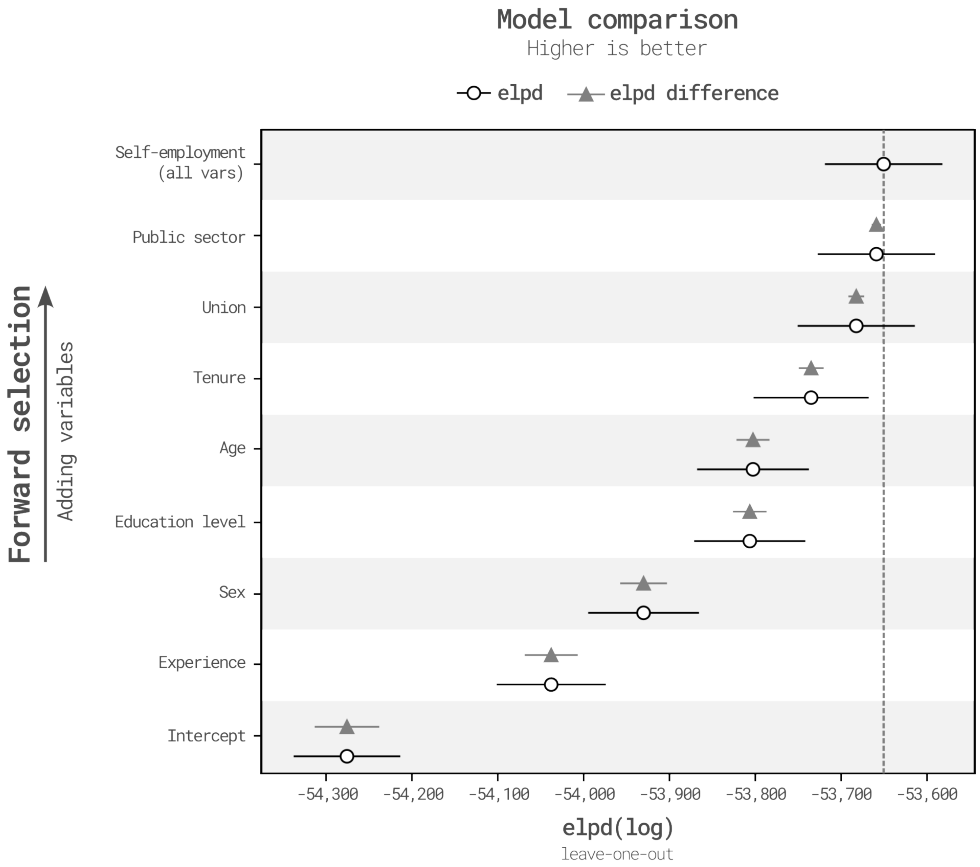

Figure 5.7: Forward variable selection

Starting from the intercept (avg_salary), each variable is added to the model and the *elpd* measure is calculated. Although adding each variable improves the model performance, the variables with the most significant effect on the target variable are experience, sex, education, and employment sector. These results are in line with the theoretical models and the literature review. On the other hand, variables such as age, tenure, and self-employment add little or no information at all, so they can be dropped from the model without losing significant performance. This selection process helps to reduce the model complexity and reduces the risk of overfitting the data.

Based on this variable selection process, the optimal model contains the predictors: ***Avg_salary, Experience, Sex, Education, Union and Public Sector***. Figure 5.8

shows a subset of the posterior distributions for some industries and occupations in the final model. The chart with all model parameters is presented in Appendix B.

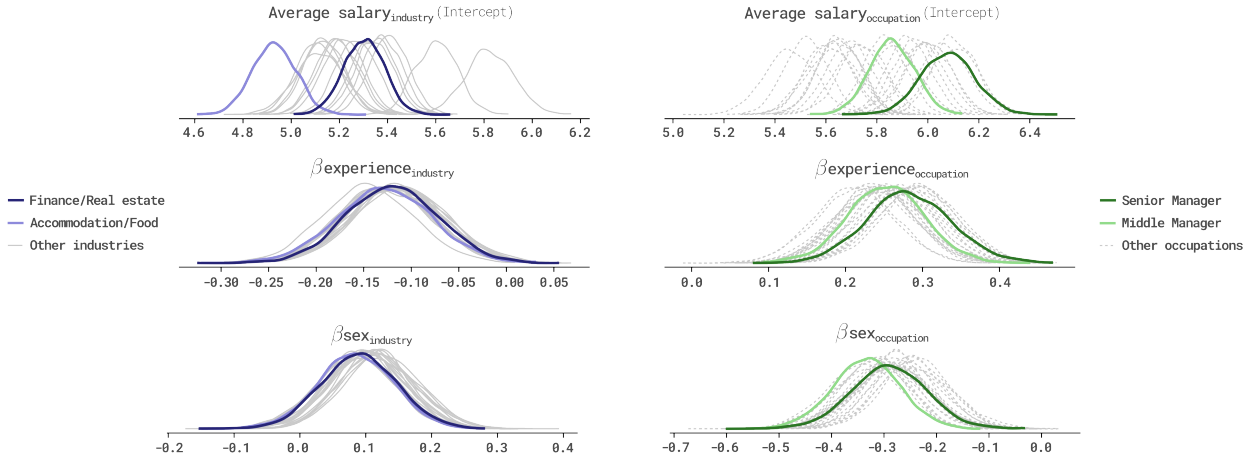

Figure 5.8: Posterior distributions for some industries and occupations in the final model

5.5 Model interpretability

The following bullets discuss the main results based on the posterior parameter distributions obtained from the Optimal model (after forward selection). It is important to highlight that although these results are in line with theory, they reflect the cultural and societal conditions from 1996 to 2007:

- Given that the Gamma GLM uses a log-link function, the relationship between the model parameters and the target variable is multiplicative[2]. Therefore, the intercept (avg_salary by industry and occupation) is interpreted as the base salary. Then, each parameter increases or reduces by a certain percentage of this salary base according to their effect (positive or negative)[3].

- As expected, most of the salary variability comes from the industry and occupation cat-

[2] After applying the inverse of the log-link (exponential), which converts the value to the units of the data (CAD 2023)

[3] After applying the inverse of the log-link (exponential)

egory, given the hierarchical structure of the labour market. In particular, the salaries of individuals in some physically demanding occupations (e.g. Construction trades) are explained mainly by their occupation rather than by the industry. This means this occupation has a low industry-base salary but higher occupation-base salaries than the other industries and occupations.

- The gender gap is mainly explained by the negative values of the sex parameter related with occupation. Although this parameter is positive for industry level, when considering the overall effect (considering both industry and occupation value) the difference is negative for all industries and occupations.
- Based on the forward selection results and the magnitude of the parameters, education level is the predictor that contributes the most to the variability explanation. As supported by the human capital theory, education is positively correlated with salaries across all industries and occupations.
- Experience is a better predictor of salaries than job tenure because the latter is attached to the current job. Conversely, experience can be seen as a proxy for the skill set of an individual. Therefore, it provides more information about the potential salary of a candidate.
- According to the model, public sector and union participation positively correlate with salaries. As public employees and unions can have more bargaining capacity, this could benefit the individuals in this sector and under union agreements.

5.6 Online learning: Longitudinal analysis of the estimated parameters

Labour markets are constantly evolving. Some changes are observed in short periods, usually cyclical and related to the oscillation between supply and demand. In contrast, some changes are observed in more extended periods, usually related to technological, social, or political changes. The former is more straightforward to predict, while the latter is challenging because

it usually corresponds to structural changes. Given the effect of the long-term changes in the labour market structure, the evolution of the model parameters through time becomes relevant for measuring the model validity and robustness to structural changes.

The model performance can be assessed through time by comparing the estimated parameters using the whole dataset versus the estimated parameters using a sequential approach. This sequential estimation is aligned with the Bayesian inference framework because the posterior belief on the current period can become the prior belief for the next period. This approach is called *Online learning*.

Using the hierarchical model after the variable selection, Figure 5.9 to Figure 5.11 compare the evolution of the posterior distribution for some model parameters using Online Learning and the same parameter posterior distributions estimated with the whole dataset.

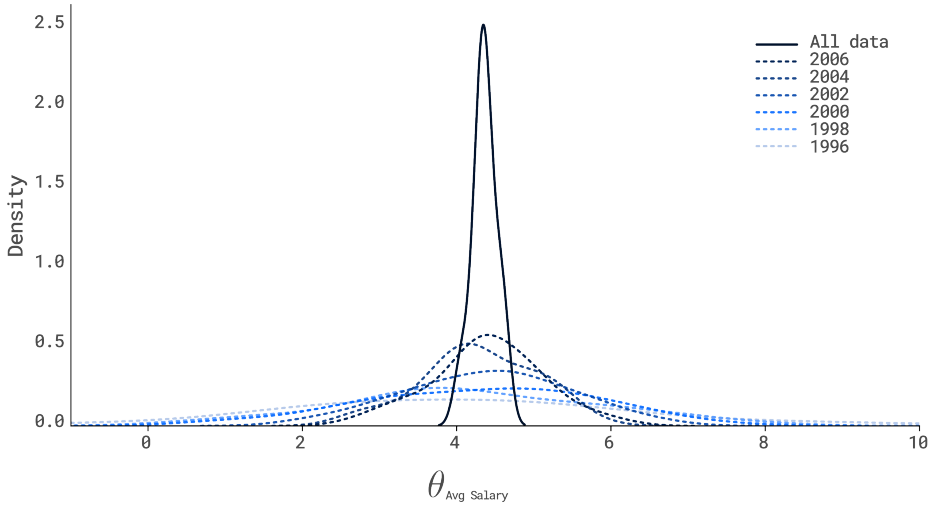

Figure 5.9: Longitudinal analysis for model parameter - Avg. Salary (Intercept)

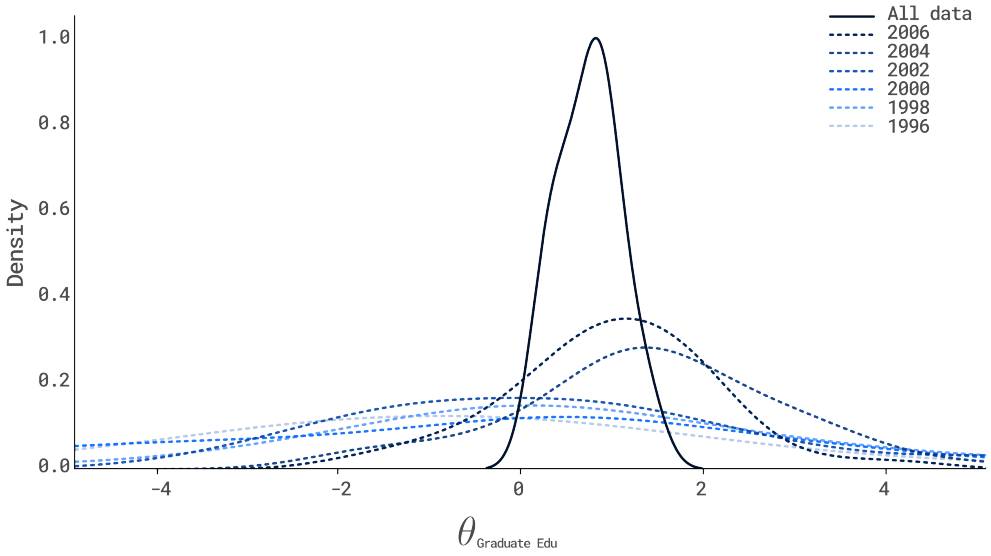

Figure 5.10: Longitudinal analysis for model parameter - Graduate Education Level

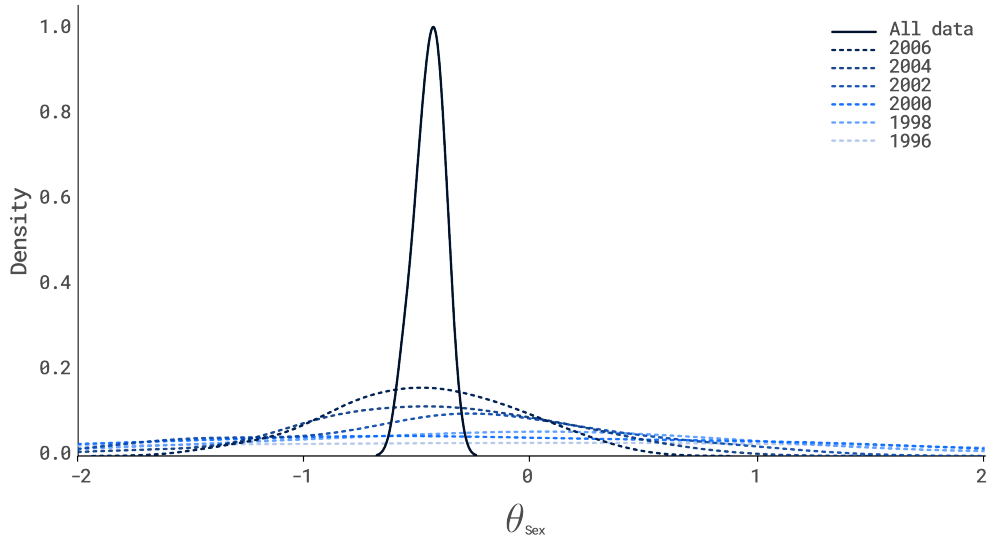

Figure 5.11: Longitudinal analysis for model parameter - Sex

The longitudinal analysis shows how the posterior distribution of each parameter is updated sequentially in light of new data. Starting from 1996, the priors are located at 0 for all model parameters, but the posteriors move towards the actual value each year. In the case of Average salary and Graduate level, they move positively, whereas in the case of Sex, they move negatively. As more data are available, the posterior distribution starts peaking around the true value, implying that the model is getting more confident in the parameter estimation. When comparing the posteriors using the whole dataset versus the sequential update (online learning), there is a difference in the density value. However, the peak of both scenarios is at the same point, which demonstrates that the Online learning setting produces reasonable estimates with a smaller dataset.

Therefore, the Online Learning technique also allows an easy and natural way to update the model parameters because it resembles how data are generated. Also, the model is more robust to structural changes because it combines the past and current structure with the trend changes in the labour market. It requires less data and computational power because the estimation task is performed gradually.

5.7 Model validation

While the in-sample metrics provide a general idea of how well the model fits the existing data, the out-of-sample performance measures the model's capability to generalize in the light of new data. That generalization determines whether the model captures the complex relationships between the model variables and ensures that the model does not "memorize" the noise in the data. Therefore, the out-of-sample evaluation assess the model's performance in a real setting.

A graphical comparison between the true and the estimated salary distribution for the validation dataset is carried out to visually assess the potential use of the predicted distributions in a simulation setting, such as the labour market module in the ILUTE framework. This validation is performed in two levels: the aggregated that ensures the model is correctly representing salaries within different clusters observed in the data (e.g. salary distribution by industry, occupation, gender, and education level, among others) and the disaggregated

that ensures the salary is being correctly predicted for each worker given their attributes.

5.7.1 Aggregated level

The following charts show a comparison between the true salary distribution and the predicted salary distribution for different group variables, being industry and occupation, the most relevant groups given the hierarchical structure of the labour market. From these figures, it is observed that the model adequately represents the salary across all group variables regarding their shape and values.

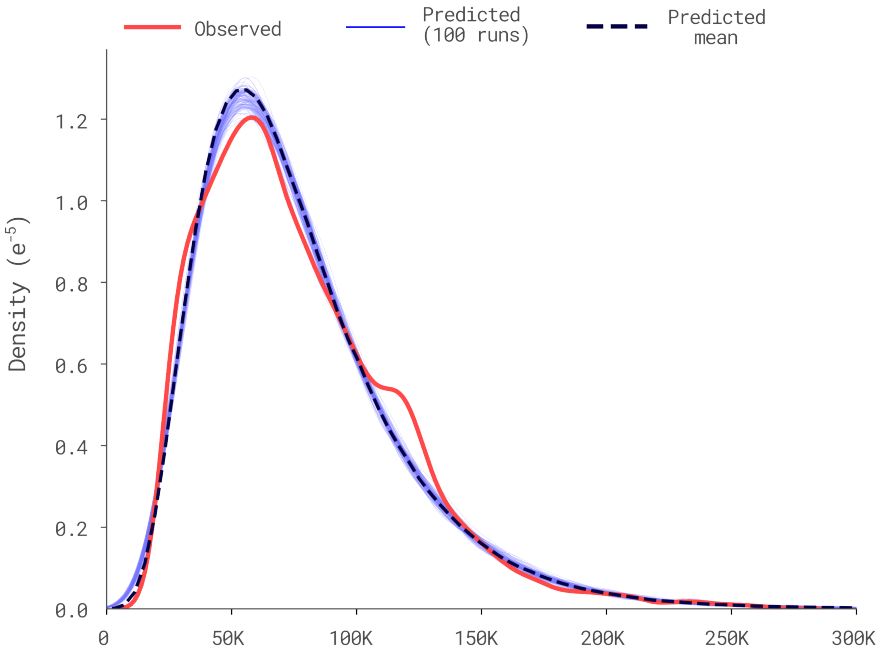

Figure 5.12: Observed and predicted salary distribution for all individuals in the validation set

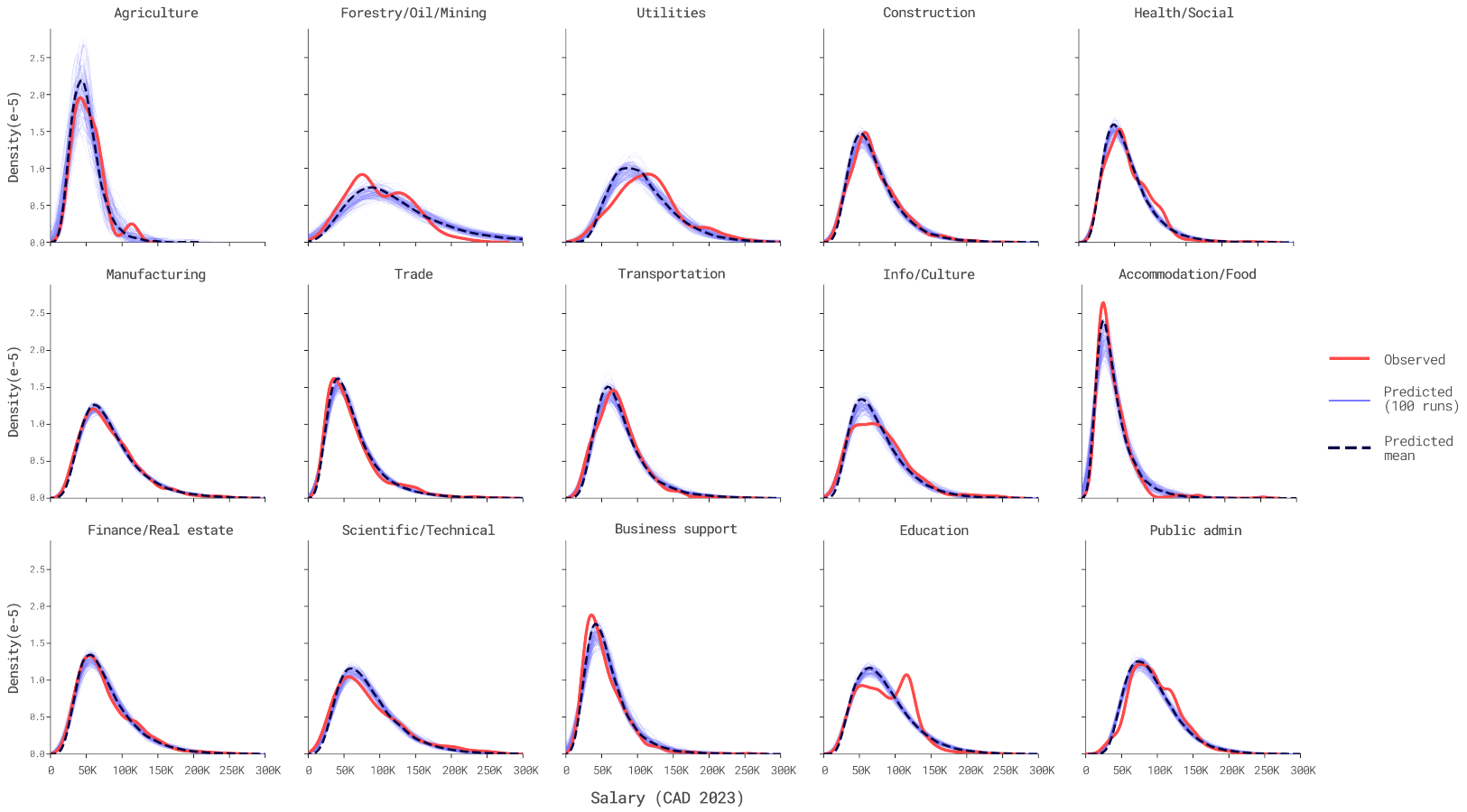

Figure 5.13: Observed and predicted salary distribution by Industry

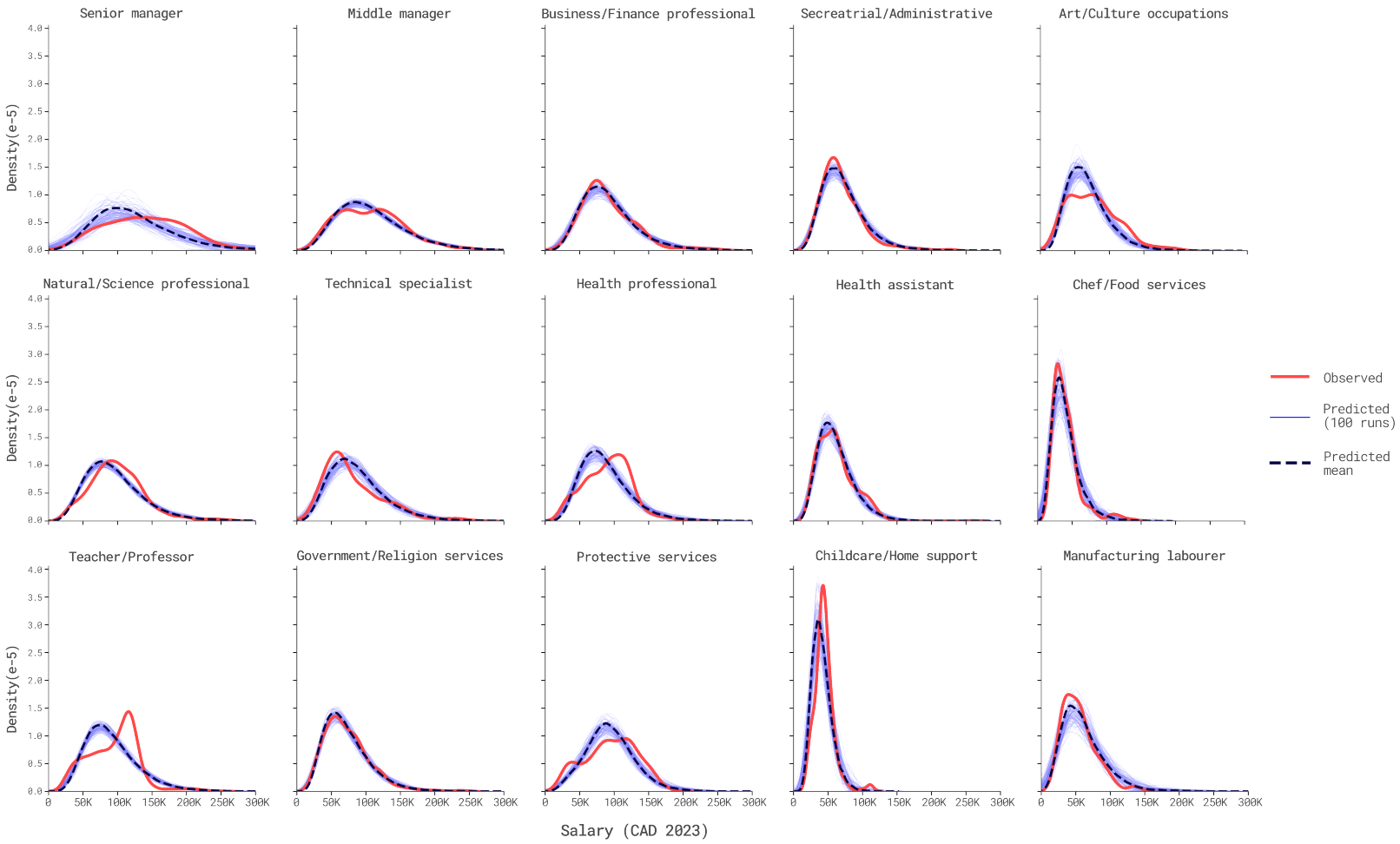

Figure 5.14: Observed and predicted salary distribution by Occupation

5.7.2 Disaggregated level

For the disaggregated level, Figure 5.15 compares the following elements for a random set of workers.

- The observed salary: salary reported by the worker in the SLID.
- The predicted salary: the expected value of the posterior salary distribution using the final model.
- The posterior salary distribution is the estimated salary distribution for that worker given their attributes.
- The posterior salary distribution of similar workers: This is the reference case to compare with the previous distribution (posterior salary distribution) and represents the average salary distribution of workers with the same attributes (variables used in the model estimation). This distribution also allows us to identify whether the observed salary is an outlier compared to workers with the same characteristics.

The comparison between these four elements provides a detailed perspective of the model's ability to represent data at the disaggregated level. The comparison between the observed salary and the posterior salary distribution of similar workers illustrates the stochastic nature of salaries. Given the size of the validation dataset, it is unfeasible to report this chart for all workers. However, the GitHub repository of this thesis provides a Jupyter notebook that allows to explore and create this chart for any worker in the dataset.

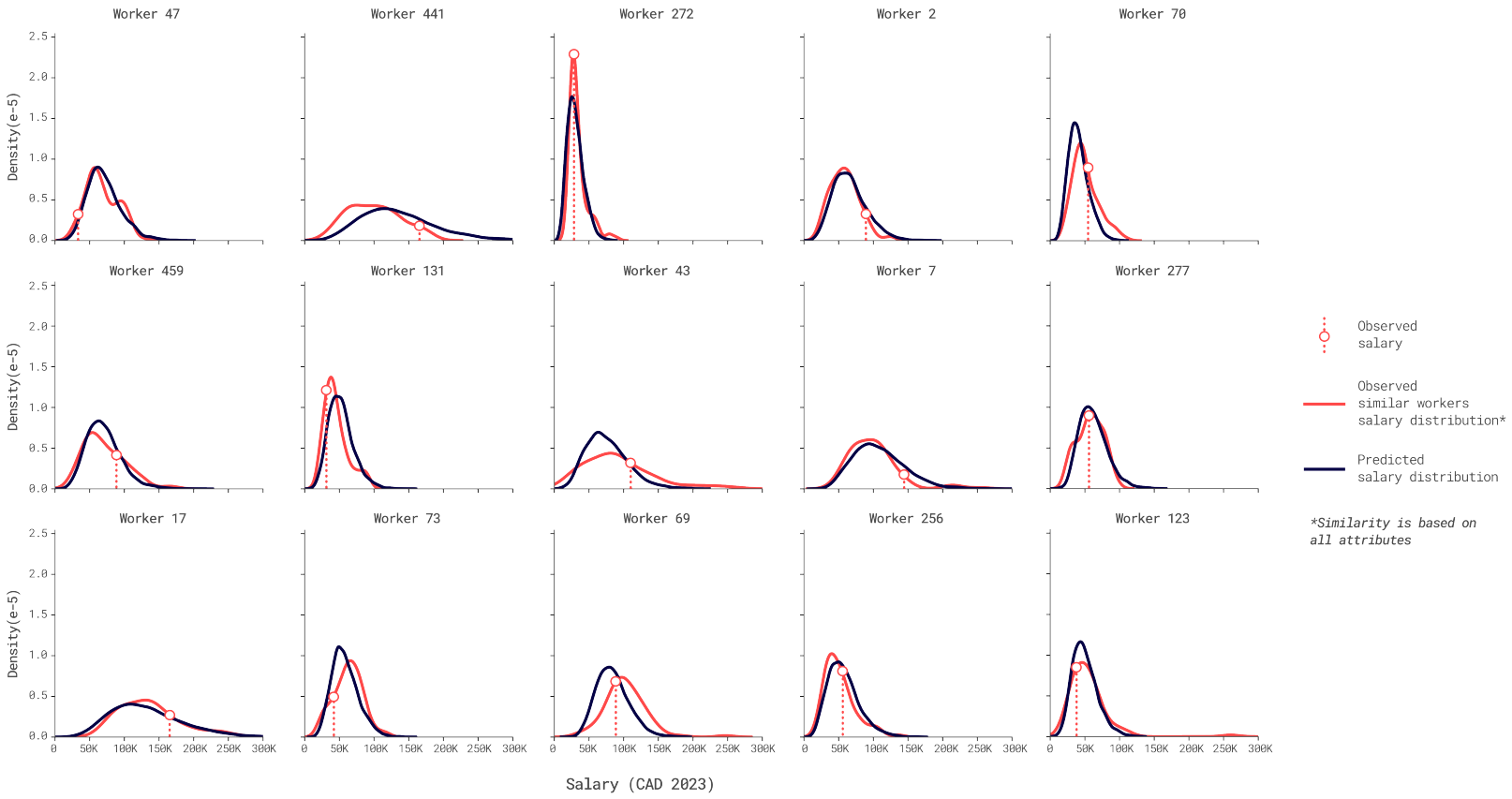

Figure 5.15: Disaggregated analysis of salary distribution for 15 random selected workers within the validation dataset

Chapter 6

Integration of the proposed model within the ILUTE framework

This chapter outlines the ILUTE labour market module changes necessary to integrate the proposed salary model. Moreover, it evaluates the proposed model relative to the current implementation using several performance criteria. Although this integration is out of the scope of this thesis, it provides an informed opinion about its potential performance within that framework.

6.1 Integration of the proposed salary model with the existing implementation

Given the difference between the outputs in the existent wage model and the proposed salary model, it is necessary to modify the salary representation within the ILUTE framework. According to Harmon (2013), salaries are represented in ILUTE as an attribute within the Person class. This attribute corresponds to the numerical value *Wage* that stores the hourly wage of each agent. However, in the proposed salary model, the prediction result is a list of values representing the annual salary distribution for that person at that time step. Therefore,

the *Wage* attribute in the Person class needs to be replaced by two attributes:

- ***Current_salary*** stores the salary value resulting from the job matching process known as *FinalOffer* (Harmon, 2013). This value plays the same role as the *Wage* attribute within the ABM but differs in its units.
- ***Reference_salary*** stores the list of values from the proposed salary model and represents the market salary distribution for an individual with such characteristics at the current time step. The size of this list corresponds to the number of samples defined in the prediction step. As shown in Figure 6.1, more samples will improve the precision but increase the memory requirements for each run and the running time for salary calculations (e.g. calculate the expected salary value for each person or compare salary distribution between individuals). Then, the right balance between sample size and precision should be decided according to the available computational resources and the running times.

Then, the Job-matching procedure remains the same, except that the *MinimumWageAccepted*, Worker's utility $\Delta V(best)$, and all premium and discount percentages[1] are calculated with respect to the expected value of the *Reference_salary*.

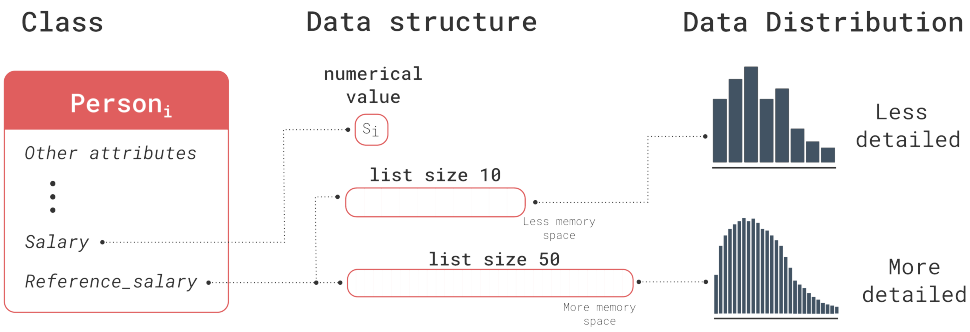

Figure 6.1: Proposed salary representation in ILUTE.

[1] *MarketAdjustmentFactor*, the premium for being employed, the discount for entering the Workforce, and the premium or discount for each matching attempt for both firm and worker: *JobApplicationAttempts* and *FailedRecruitingAttempts*

6.2 Model comparison: proposed vs existent salary model

As the units of the target variable in the existing and proposed model differ, a direct comparison between performance metrics might not be the best approach to compare them. Hence, the following list compares both models based on different criteria such as performance, computation efficiency, and suitability within the ILUTE framework.

- **Model capability**: In the original implementation, Hain (2010) defined a series of interaction terms to improve the model's performance. These interaction terms capture the complex relationships and dependencies between the variables, but this process is performed manually. In the proposed model, the hierarchical structure captures these relationships and dependencies without using any manual process. Therefore, the proposed model is a better approach for representing these dependencies between variables without risking the introduction of some bias in the model.
- **Model Interpretability**: Although introducing predictor variables could increase model performance, it directly reduces the model's interpretability. This would not be an issue in some settings where prediction accuracy is the aim. However, interpretability becomes very important when the model objective is to support public policy decisions, such as in ILUTE. Therefore, the proposed model contains ten variables with no explicit interaction terms, which allows a more straightforward interpretation of the model results and parameter meaning.
- **Model complexity**: There is a direct relationship between the number of predictors (variables and interaction terms) and the model complexity. The higher the model complexity, the higher the risk of overfitting the data, which decreases the model's ability to generalize on new data and reduces the model's robustness to structural changes in the future labour market. The shrinkage effect observed in Section 5.4.1 and using a holdout dataset to measure the real performance Section 5.7, reduces the risk of overfitting and improves the model's generalization capability.
- **Model purpose and usability**: The labour market usually involves random varia-

tions and uncertainty in its processes and outputs. For instance, two individuals with the same attributes can have salary differences explained by random factors such as belonging to a company that pays better or having better negotiation skills. Although some systematic components explain most salary distribution, some can only be explained by introducing randomness in the modelling process. This randomness makes salary prediction a stochastic process in nature. However, the original implementation of the wage model produces a deterministic output, which means that a set of initial conditions will always produce the same outputs. Given the simulation nature of the ILUTE framework, the proposed model seems to be a better approach to introduce this random component in the prediction by producing distributions instead of point estimates. In particular, these distributions might improve the job-matching algorithm and the overall ABM process proposed by Harmon (2013). On the other hand, as the ILUTE framework is supposed to be used in a public policy evaluation setting, a what-if evaluation could benefit more from a stochastic perspective than a deterministic one by accounting for the uncertainty in the estimation of the results both at the aggregate or disaggregated level.

- **Model robustness and flexibility**: In the long term, the labour market evolves by introducing or eliminating some industries and occupations, adjusting the market to political and cultural changes, and reflecting technological changes in the salaries received by workers in a specific sector, among others. The hierarchical structure of the proposed model eases the introduction of these changes, as new industries and occupations generally share some characteristics with the existing industries. Hence, the information transfer between different levels in the proposed model produces better prediction under these changing scenarios. Even if this first guess is not accurate, the online learning structure in the Bayesian approach helps to update these estimates in the light of new data. Conversely, the original model must be entirely estimated when new data is available or when there are significant changes in the labour market structure. This estimation could require a significant effort and produce contradictory results between each estimation because there is no way to use the previous parameters in the new estimation process.

- **Model inference efficiency**: The advantages of the proposed model discussed in the last bullets come with a cost. The complexity of the estimation process increases considerably in the proposed model because new processes, such as the prior definition and the sampling process, are additional steps compared with the simplicity of the ordinary least squares (OLS) of the existent model. Given the probabilistic nature of the sampling process, this is a slow and computationally inefficient process because a percentage of the samples are discarded, which means that a portion of the allocated computational power is wasted. Despite recent improvements in efficient alternatives to MCMC, this process is still slow compared to the original approach (OLS). A way to mitigate this issue is to define the priors that represent the observed phenomena carefully. This helps the model converge more efficiently and reduces the number of discarded samples.

- **Model memory usage**: Although the proposed model provides more information than the original approach, this advantage could become a disadvantage because this information needs to be stored and handled adequately. While the memory constraint is an issue that is becoming less important as powerful hardware is getting cheaper (Moore's Law), it is still relevant when dealing with a multi-agent simulation environment. In this case, an adequate model implementation27 and a proper sample size definition can help reduce memory usage while producing helpful information.

- **Model scalability**: Although this topic applies to both approaches, it is known that MCMC sampling scales poorly as more data is added to the model. To mitigate this effect, it is crucial to define the process of updating the model parameters using approaches like the one discussed in Section 5.6. This is an advantage of the proposed model compared to the original approach because the existing model needs to be estimated from scratch in the light of new data. However, the difference in the computational cost of both approaches is better for the original model.

- **Temporal scale representation**: The existing model generates estimates at the hourly level, and the proposed model predicts salaries at the annual scale. While the proposed model can produce better estimates of the individual annual income, the

hourly approach of the existent model provides more flexibility to model different temporal scales as hourly salary can be aggregated into several representations such as daily, monthly or annually. However, as pointed out in Harmon's work, some assumptions about the number of work hours per week can induce estimation errors, affecting the original model's reliability and performance.

Chapter 7

Conclusions

The hierarchical structure in labour markets defines many attributes of both the firms and the individuals. According to the findings in this thesis, this structure is closely related with the observed wage differentials. Therefore, the inclusion of this structure in the modelling process will improve significantly the results and accuracy of the model. However, this structure also needs to consider the relationships between the different levels. The hierarchical or multilevel model is an approach that meets these requirements because it represents the hierarchical structure while also models the complex relationships between the multiple levels by sharing information between them.

Given the stochastic nature of simulations, the use of a random component seems to be more flexible and realistic than the use of point estimates to represent salaries within the labour market. As salaries are positive a right skewed, the application of the regular linear regression model could be misrepresenting the random component of salary definition. As suggested by the literature, the use of the Gamma distribution in conjunction with a systematic component improves the model performance and allows to capture the particularities of the observed salaries.

Besides the model structure, the application of the Bayesian inference framework allows to extract more information from the data by producing probability distributions instead of

point estimates. This is more evident when modelling out-of-sample salary distributions at both the aggregated and disaggregated level with high accuracy and a good bias-variance balance, which demonstrates its potential to produce robust estimations. However, these advantages come with a cost in terms of computational power.

7.1 Future work

Based on the results presented in Chapter 5 and Chapter 6, the proposed model needs to be tested within the Harmon's AMB implementation of the labour market in ILUTE. Although Chapter 6 compares the methodologies of the existent and proposed model, applying the last one in a simulation setting will provide evidence whether the use of salary probability distributions outperforms the use of point estimates.

As the hierarchical model is also applicable to spatial problems, the proposed model could benefit from including job location as a predictor at the industry level. This approach will improve the representation of the decision-making process within the job matching process in ILUTE and provide a more realistic.

The proposed model is not considering any variation between the model's parameters. However, a possible way to improve the predictive accuracy could be to model the covariance between the predictors by setting a prior over the correlation matrix using the LJK distribution. This distribution provides a way to define a uniform distribution of all possible positive defined correlation matrices. With this approach, some labour market distributions such as the gender and level of education can be better represented for some specific industries and occupations.

On the other hand, although the SLID dataset is a detailed data source, future efforts should be focused on using more updated datasets to validate the effects of recent changes in the labour force in the estimated models. In this matter, Statistics Canada has an interesting set of statistical programs but most of them require an application process to access this data. Under the light of new data, the proposed model is expected to benefit from the online learning scheme, in which the model can be easily updated under the Bayesian framework.

This update just consider that the results of this thesis are the priors to be updated using the new data.

Another important aspect to consider when evaluating the possible effects of current and future changes in the labour market is the increment of salaries over time. Given the stability observed in most of the salaries during the period of study, the autoregressive nature of the time series could be obviated. However, as the labour markets are continuously changing this behaviour can be different in the future. Therefore, this is something to consider in future updates of this model.

Bibliography

Anderson, J. M. (1997). Models for retirement policy analysis.

Angrist, J. D. and Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size.

Baker, M. and Benjamin, D. (1999). Early retirement provisions and the labor force behavior of older men: Evidence from canada. *Journal of Labor Economics*, 17:724–756.

Benjamin, D., Gunderson, M., Lemieux, T., Riddell, C., and Schile, T. (2021). *Labour market economics : theory, evidence, and policy in Canada*. McGraw-Hill Ryerson, 9th edition edition.

Borjas, G. J. (2020). *Labor economics*. McGraw-Hill Education, 8th edition edition.

Devore, J. (2016). *Probability and statistics for engineering and the Sciences*. Cengage Learning.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Texts in statistical science. CRC Press Boca Raton, Boca Raton, third edition edition.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.

Gunderson, M. (1979). Earnings differentials between the public and private sectors. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 12:228–242.

Hain, M. D. L. (2010). Labour market model of the greater toronto and hamilton area for integration within the integrated land use, transportation, environment modelling system.

Hamilton, B. (2000). Does entrepreneurship pay? an empirical analysis of the returns to self-employment. *Journal of Political Economy*, 108:604–631.

Harmon, A. (2013). A microsimulated industrial and occupation-based labour market model for use in the integrated land use, transportation, environment (ilute) modelling system.

Harmon, A. and Miller, E. J. (2018). Overview of a labour market microsimulation model. *Procedia Computer Science*, 130:172–179.

Harmon, A. and Miller, E. J. (2020). Microsimulating labour market job-worker matching. *Journal of Ambient Intelligence and Humanized Computing*, 11:993–1006.

Kaufman, B. and Hotchkiss, J. L. (2003). *The Economics of Labor Markets*, volume 1. South-Western College Pub.

Keegan, M. (2007). Modelling the workers of tomorrow: the appsim dynamic microsimulation model.

Lambert, B. (2018). *A student's guide to Bayesian statistics*. SAGE Los Angeles, Los Angeles.

Lewis, H. G. (1986). Union relative wage effects. *Handbook of Labor Economics*, 2:1139–1181.

McCurdy, T. E., Brown, D., Gould, W., Michael, R., Pencavel, J., Sumner, D., and MaCurdy, T. E. (1980). An empirical model of labour supply in a life cycle setting.

Mcdonald, I. M. and Solow, R. M. (1992). Wage bargaining and employment. *Economic Models of Trade Unions*, pages 85–104.

McElreath, R. (2016). *Statistical rethinking : a Bayesian course with examples in R and Stan*. Texts in statistical science. CRC Press/Taylor & Francis Group Boca Raton, Boca Raton.

Miller, E. J. and Vaughan, J. (2021). Place of residence-place of work modelling: Issues & options.

Neal, R. M. (2012). Mcmc using hamiltonian dynamics.

Nielsen, S. F. (2010). Generalized linear models for insurance data. *Journal of Applied Statistics*, 37:703. doi: 10.1080/02664760902811571.

Phan, D., Pradhan, N., and Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in numpyro.

Salvini, P. and Miller, E. J. (2005). Ilute: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and Spatial Economics*, 5:217–234.

Smith, A. (1910). *The Wealth of Nations*, volume 1. London : J.M. Dent & Sons ; New York : E.P. Dutton.

Smith, S. (2003). *Labour Economics*. Taylor & Francis Group, 2nd edition edition.

StatisticsCanada (1998). Overview of dynacan a full-fledged canadian actuarial stochastic model designed for the fiscal and policy analysis of social security schemes.

StatisticsCanada (2012). Survey of labour and income dynamics (slid) - a 2011 survey overview.

StatisticsCanada (2013). The lifepaths microsimulation model: An overview.

Sullivan, P. (2010). A dynamic analysis of educational attainment, occupational choices, and job search.

Tikanmäki, H. and Lappo, S. (2020). Elsi: The finnish pension microsimulation model.

Vehtari, A., Gelman, A., and Gabry, J. (2015). Practical bayesian model evaluation using leave-one-out cross-validation and waic.

Appendix A

Posterior distributions

A.1 Posterior distributions for the pooled model

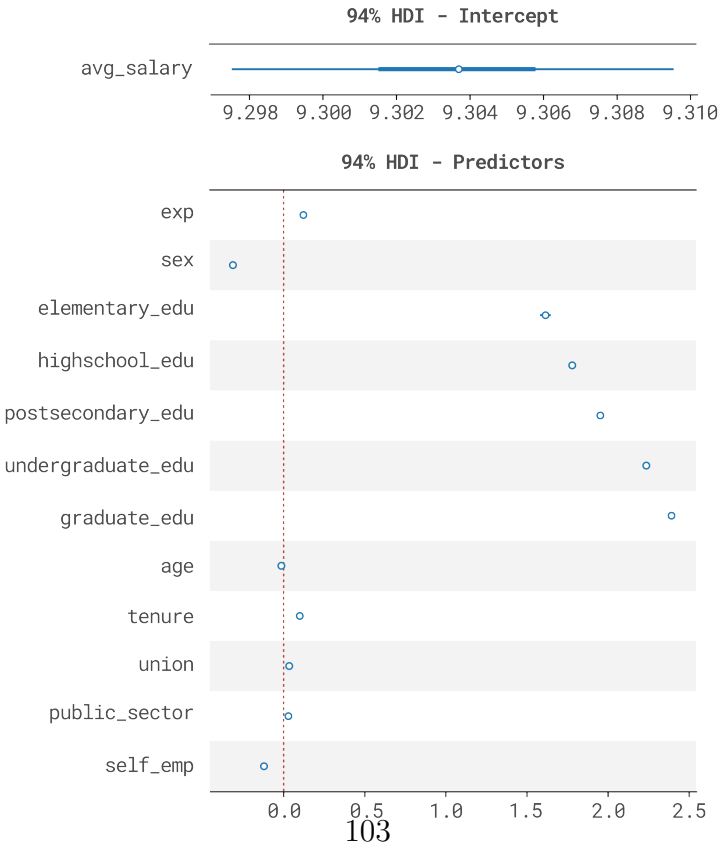

Figure A.1: Posterior distributions of model parameters for the pooled model.

A.2 Posterior distributions for the no-pooled model

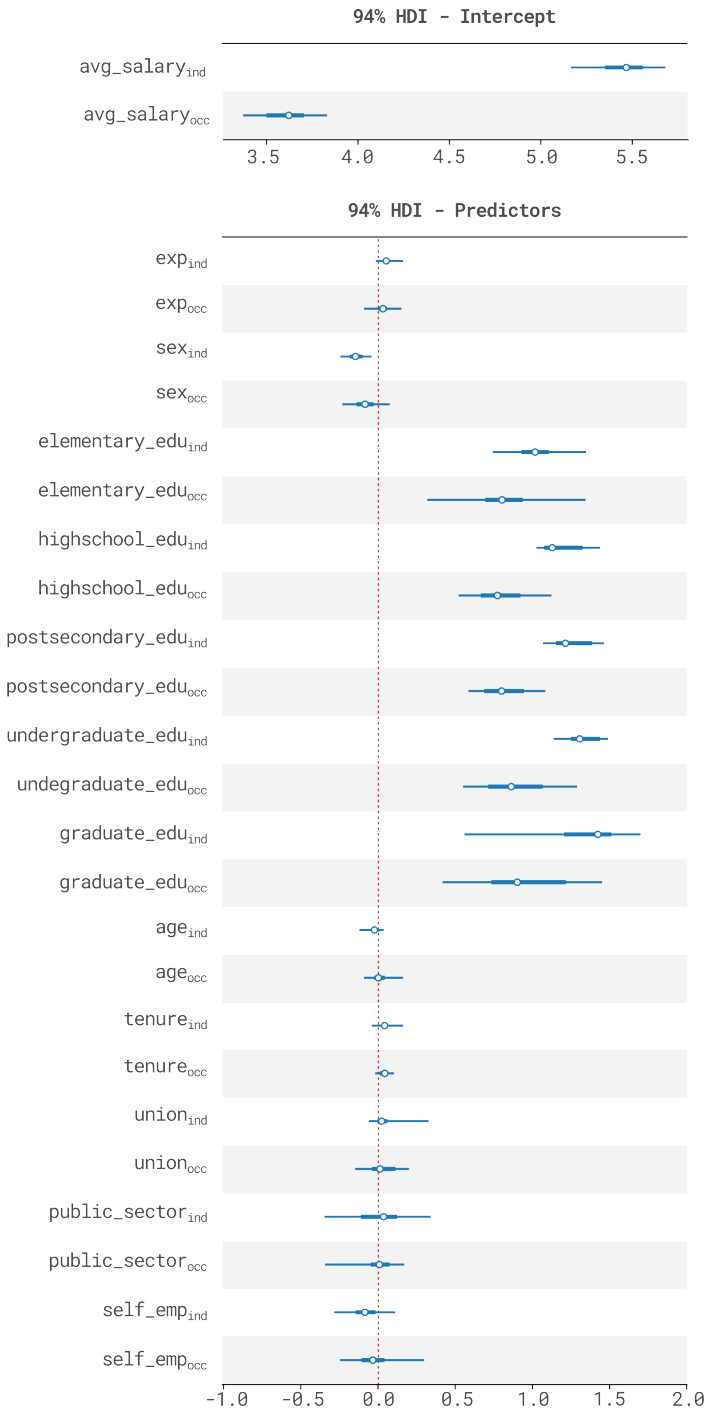

Figure A.2: Posterior distributions of model parameters for the no-pooled model (*All categories aggregated).

A.3 Posterior distributions for the hierarchical model

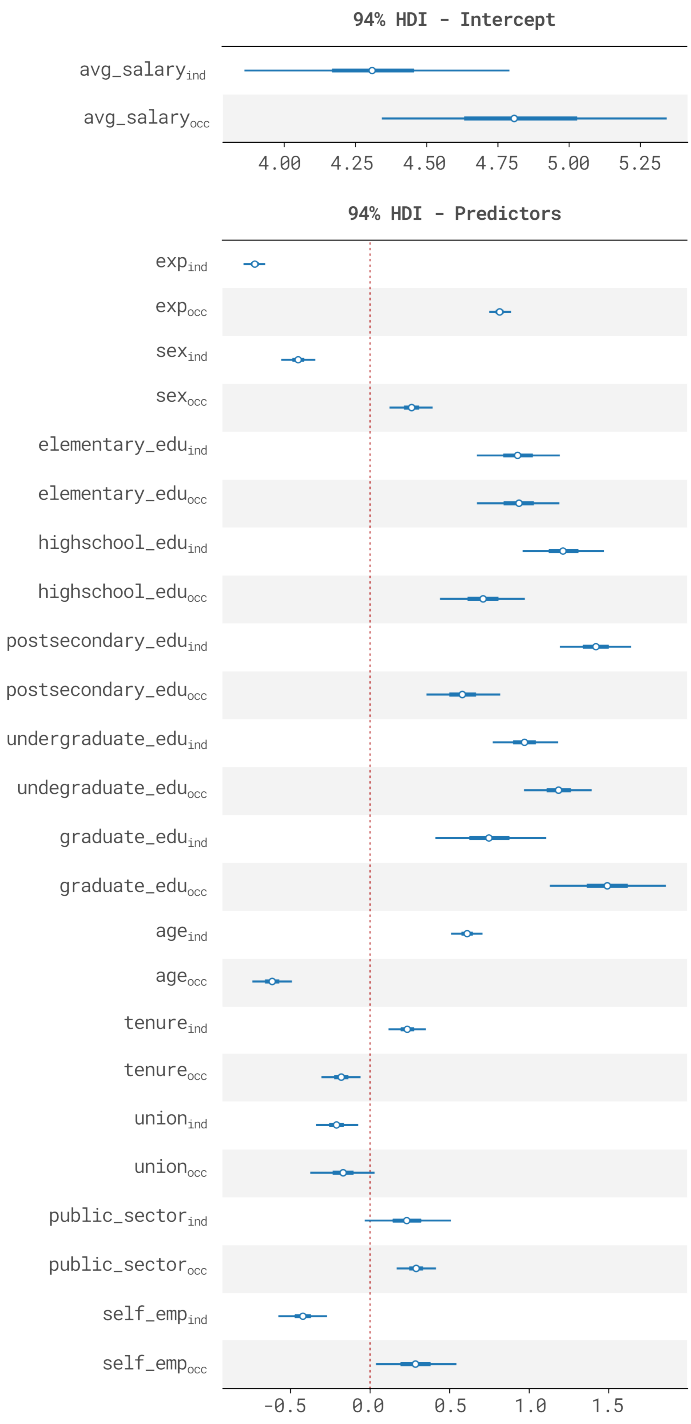

Figure A.3: Posterior distributions of model parameters for the hierarchical model (*All categories aggregated).

Appendix B

Posterior distributions for final model

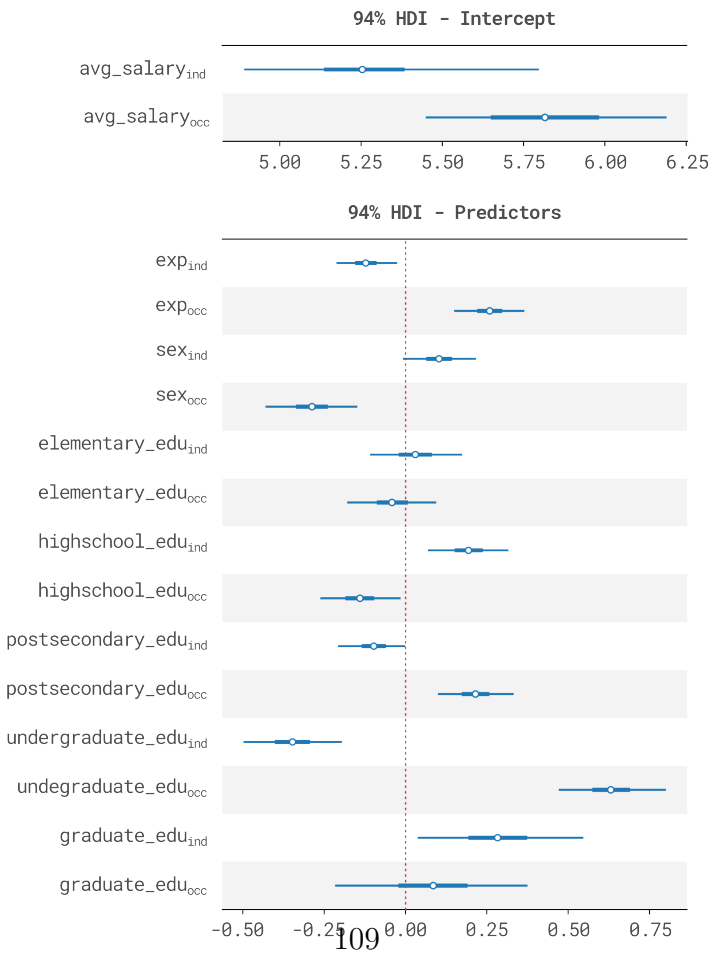

Figure B.1: Posterior distributions of model parameters for the final model (*All categories aggregated).