Predicting salaries in the GTA

A Bayesian approach

Andrés Castillo

Supervisor: Prof. Eric Miller
Department of Civil and Mineral Engineering
University of Toronto
2023

Outline

1. **Research motivation**
2. **Modelling labour markets**
   - Labour market components
   - ILUTE framework
3. **Predicting salaries**
   - Data sources
   - Model structure
   - Model specification
   - Variable selection
4. **Model validation and results**
   - Estimated parameters
   - Aggregated level
   - Disaggregated level
5. **Conclusion and future work**

Research motivation

Why we talk about labour markets in transportation?

Research motivation

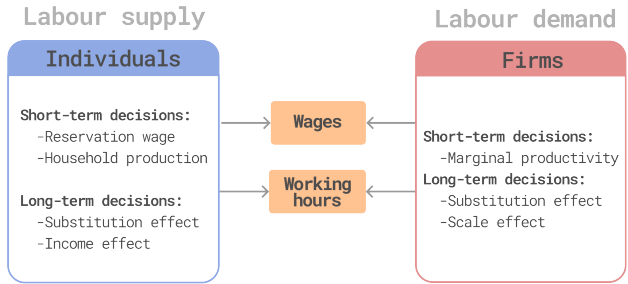
- Transport models have grown in complexity but some inputs are still considered **exogenous**.
- Home-Based Work is the **second most frequent trip purpose** in the GTA (33.8% - TTS, 2016).
- Place of residence, place of work, household income, and auto ownership are **directly or indirectly related** to outcomes of the labour market.

Modelling labour markets

How we integrate labour markets in transportation models?

Modelling labour markets
Labour market components

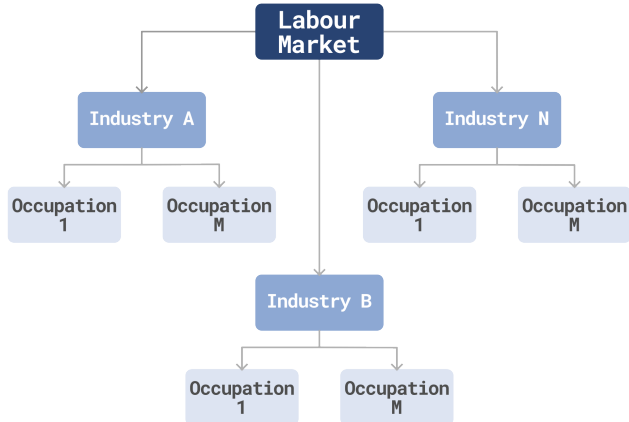
- Wages **facilitate the interaction** between agents in the labour market
- Wages **allocate labour** to the most efficient use
  - Industries
  - Occupations
  - Regions



General overview of the Labour market components

Modelling labour markets
Structure

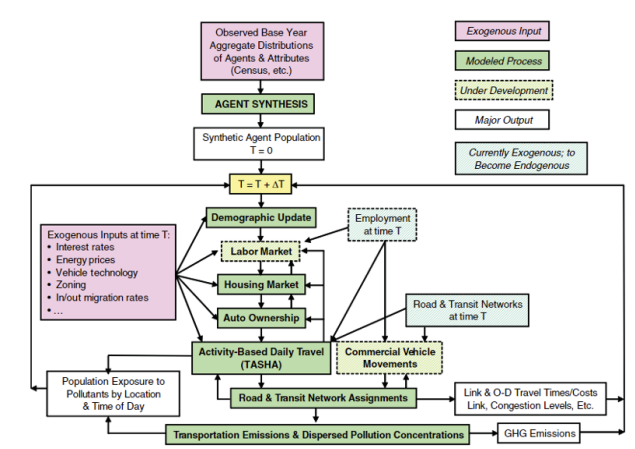
- Labour markets are organized in a **hierarchical** structure
  - Industries
  - Occupations
- This structure contributes to the **wage differentials**.



Hierarchical structure of the labour market

Modelling labour markets
ILUTE framework

- Hain (2010) proposed a **hourly wage model** and a **transition model**.
- Harmon (2013) used these models and built an ABM to simulate:
  - Worker's state update
  - Job creation/deletion
  - Job search and matching

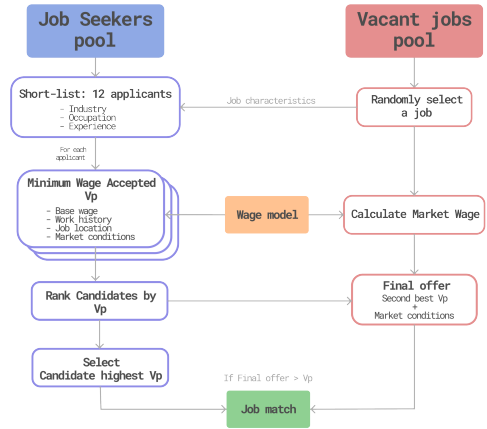


ILUTE flowchart (Miller et al, 2021).

Modelling labour markets
ILUTE framework

- Harmon (2013) found that:
  - The existent wage model is **underestimating salaries**.
  - The source of the problem could be related to the **Worked Hours** attribute.

The model is missing the random component of the wage distribution (Deterministic vs. Stochastic approach)



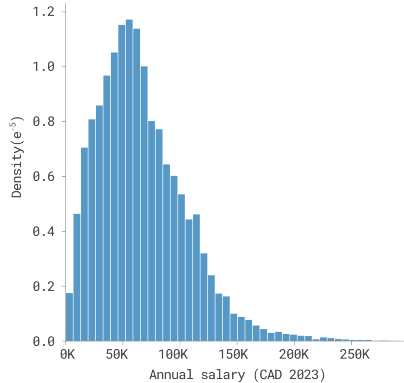
Job search and matching process in ILUTE (Adapted from Harmon, 2013).

Predicting salaries

From point estimates to probability distributions

Predicting salaries
Data sources - SLID Survey

- The **Survey of Labour and Income Dynamics (SLID) 1996 - 2011**(StatCAN):
  - National household survey - Longitudinal
  - Panel of 17K households and 34K respondents
  - Individuals are interviewed once a year for six years
- The dataset is split into two sets:
  - Training set (1996 - 2007)
  - Validation set (2008 - 2011)



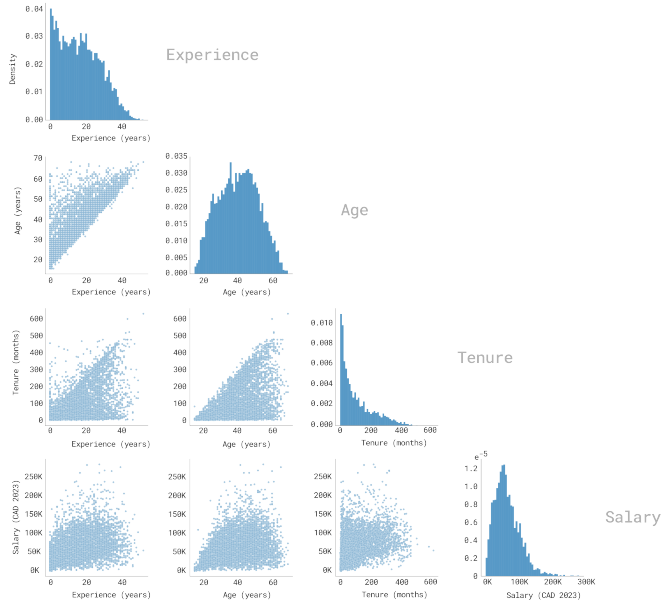
Salary distribution in the GTA 1996-2011

**Salary data is converted to real terms (CAD 2023) using the CPI.*

Predicting salaries

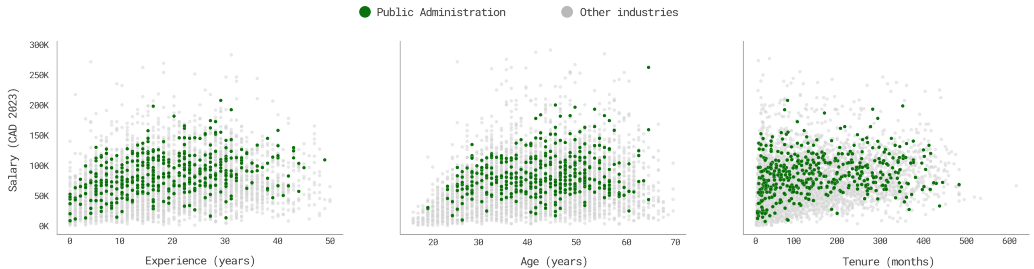
Exploratory Data Analysis

- Age, experience, and tenure are **positively correlated** with salaries.
- Weak linear relationship with **high variance**.



Salaries by experience, age, and tenure in the GTA 1996-2011

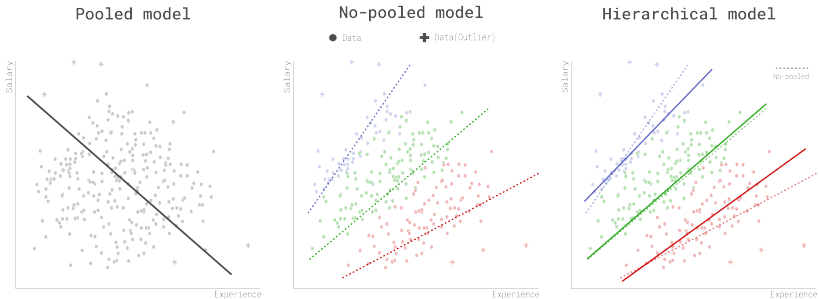
Predicting salaries
Exploratory Data Analysis



Salaries by experience, age, and tenure for Public Admin in the GTA 1996-2011

Given the hierarchical structure, the linear relationship between some predictors becomes more explicit when data is filtered by industry and occupation

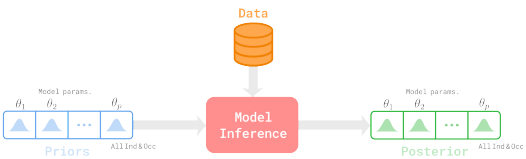
Predicting salaries
Model structure



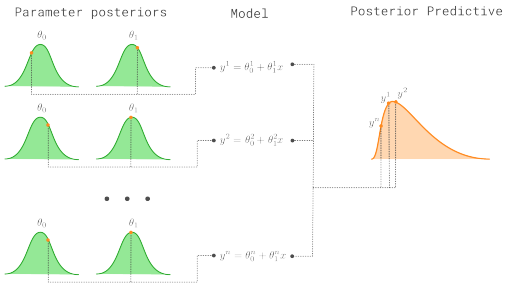
Effect of model structure on the data representation

*The Hierarchical model catptures the **data heterogeneity**, is more **robust to outliers**, and reduces the **risk of overfitting**.*

Predicting salaries
Model structure - Pooled



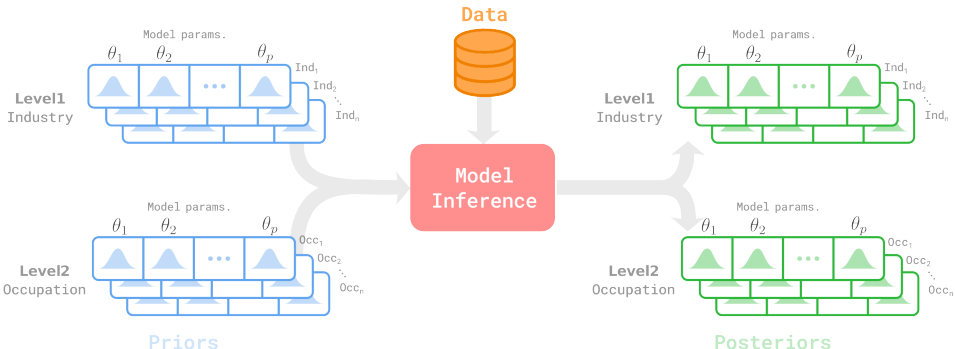
Bayesian inference for the pooled model



Predictions using the posterior distributions

Predicting salaries

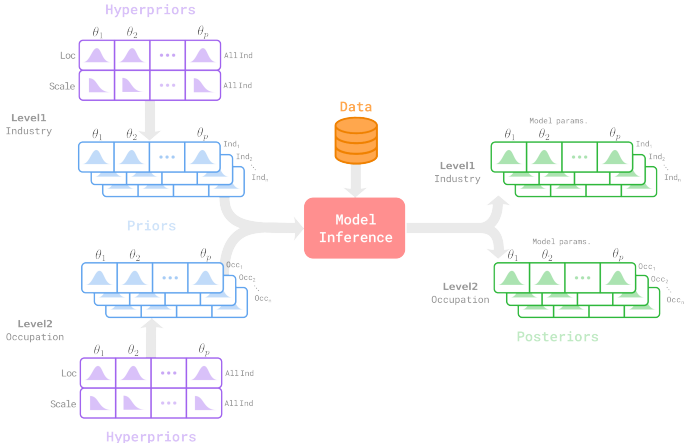
Model structure - No-Pooled



Bayesian inference for the no-pooled model

Predicting salaries

Model structure - Hierarchical



Bayesian inference for the hierarchical model

Predicting salaries

Model specification

$$loc_p^{ind} \sim Normal(0, 1)$$

$$scale_p^{ind} \sim HalfNormal(1)$$

$$loc_p^{occ} \sim Normal(0, 1)$$

$$scale_p^{occ} \sim HalfNormal(1)$$

$$\alpha \sim Uniform(0, 100)$$

$$\theta_p^{ind} \sim Normal(loc_p^{ind}, scale_p^{ind})$$

$$\theta_p^{occ} \sim Normal(loc_p^{occ}, scale_p^{occ})$$

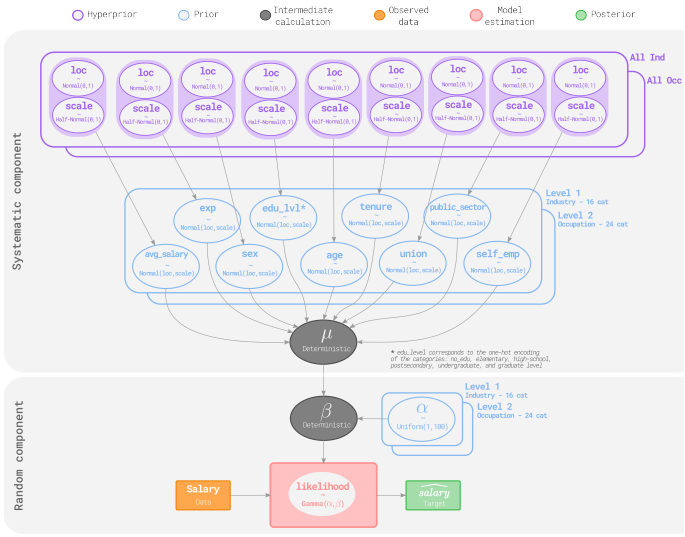
$$\eta^{ind} = \theta_0^{ind} + \theta_1^{ind} X_1 + \ldots + \theta_p^{ind} X_p$$

$$\eta^{occ} = \theta_0^{occ} + \theta_1^{occ} X_1 + \ldots + \theta_p^{occ} X_p$$

$$\mu = e^{\eta^{ind} + \eta^{occ}}$$

$$\beta = \alpha / \mu$$

$$Y \sim Gamma(\alpha, \beta)$$

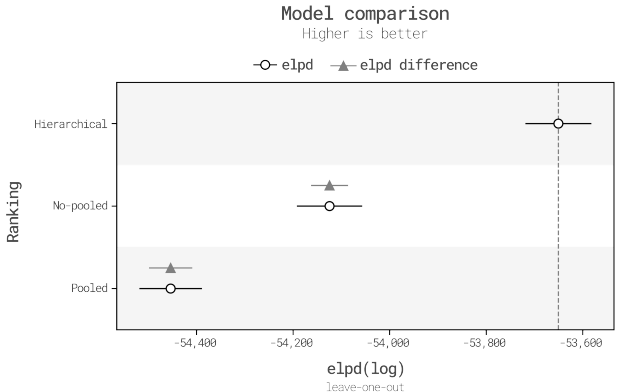


Hierarchical model graph

Predicting salaries

Model specification - Results

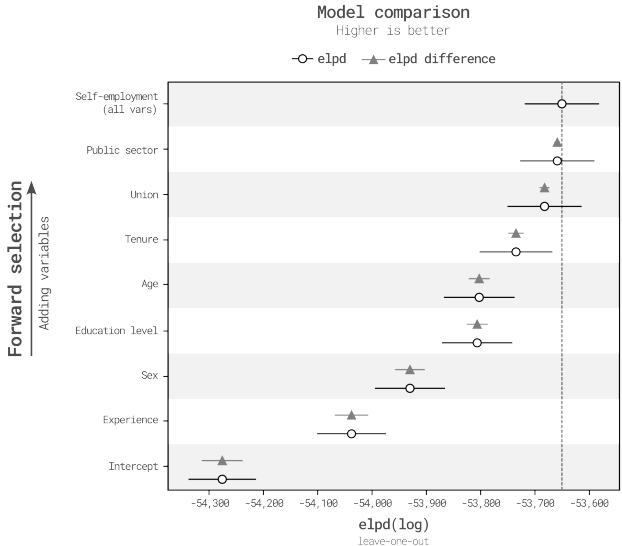
- The **expected log-pointwise density(elpd)** is an indicator of the model's predictive performance.
- It provides a proxy measure of the information loss(**KL-divergence**).



Model performance comparison of proposed structures

Predicting salaries
Variable selection

- Experience, Sex, and Education level are the **most important** predictors.
- The improvement for adding Age and Self-employment as predictor is **minimal**.



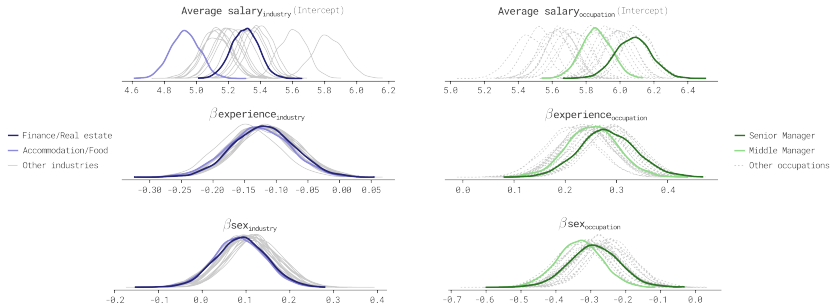
Forward variable selection process

Model validation and results

Aggregated and disaggregated level

Model validation and results
Estimated parameters

- The parameters' effect and variability is **higher** on the occupation level.
- Experience and education level have a **positive effect** in salaries.
- Overall, the **Gender gap** effect is **negative**.

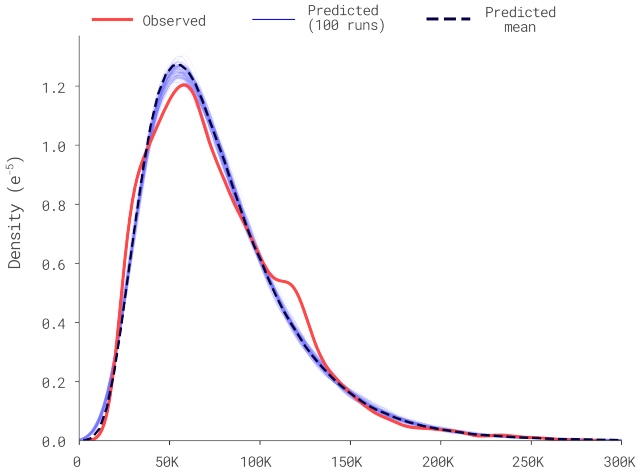


Posterior distribution for some model's parameters

Model validation and results

Aggregated level - GTA

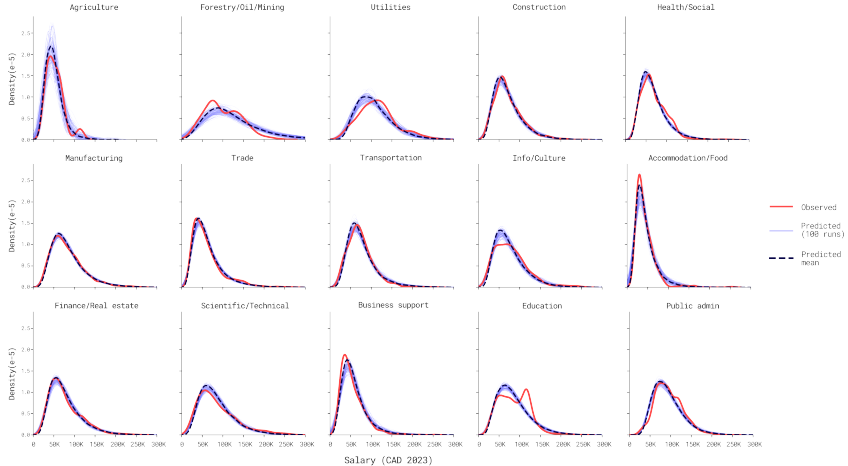
- The model validation is performed using the **validation set (2008-2011)**.
- This data was **not used** in the model estimation.
- These results provide a good measure of the **real model's predictive performance (out-of-sample)**.



Observed and predicted annual salary distribution for the GTA

Model validation and results

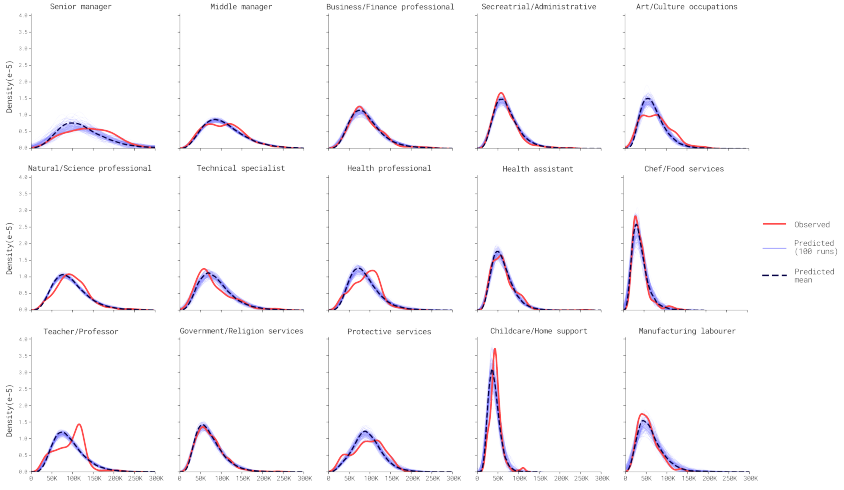
Aggregated level - Industry



Observed and predicted annual salary distribution by industry for the GTA

Model validation and results

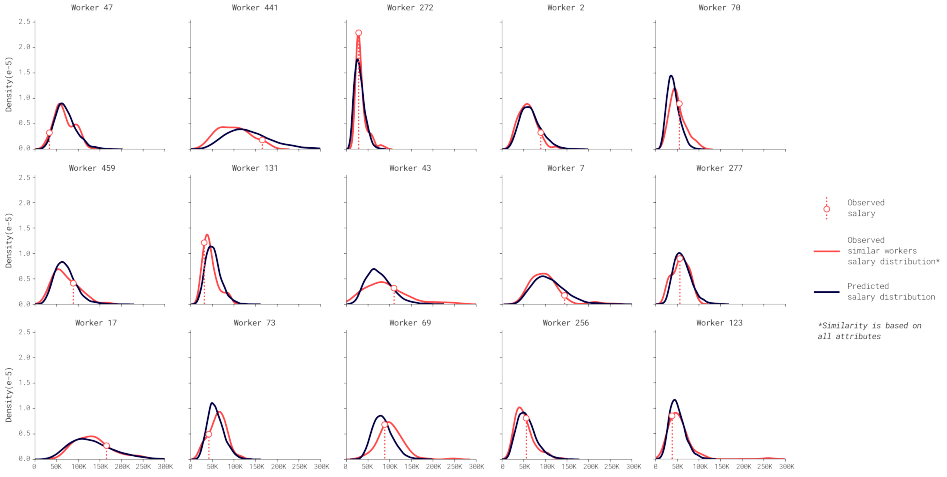
Aggregated level - Occupation



Observed and predicted annual salary distribution by occupation for the GTA

Model validation and results

Disaggregated level



Observed and predicted salaries for a group of randomly selected workers

Conclusions and future work

Conclusions

- The introduction of the Hierarchical structure in the model allows to capture complex relationships and dependencies in the data. It replaces the manual definition of interaction terms by extracting the information from the data.
- The definition of the Gamma distribution as the random component of the model allows to capture the data heterogeneity and generate better estimations.
- Labour markets have some degree of randomness. The use of probability distributions instead of point estimates produces more realistic results.
- The Bayesian thinking is suitable for modelling complex systems. It allows to update our prior knowledge based on new evidence. However, it comes with a computational cost.

Future work

- Run this model within the Harmon's ABM implementation. ***Evaluate other submodels to implement the Bayesian approach***.
- Include job location as a predictor. ***Hierarchical models are specially good in spatial applications!***.
- Model the parameters' covariance to improve the performance at the disaggregated level. ***Computationally expensive...For now!***.

Aknowledgements

To my loves: Ale, Emi, Avena. None of this would have been possible without your support

Prof. Eric Miller for his patience, guidance and support during this journey. Cities as complex systems course changed the way I see the world.

COLFUTURO for providing the partial funding for my studies.

Prof. Matthew Roorda for kindly accepting to read this thesis.

TMG group: I learned a lot from you guys. Thanks for the support and the good times!

Questions?

Thank you for your attention!