# A multilevel Bayesian framework for predicting municipal waste generation rates

Maximiliano Cubillos*, Jesper N. Wulff*, and Sanne Wøhlk*

*Department of Economics and Business Economics,
Aarhus University, Denmark

**Abstract**

Prediction of waste production is an essential part of the design and planning of waste management systems. The quality and applicability of such predictions depend heavily on model assumptions and the structure of the collected data. Ordinarily, municipal waste generation data are organized in hierarchical structures with municipal or county levels, and multilevel models can be used to generalize linear regression by directly incorporating the structure into the model. However, small amounts of data can limit the applicability of multilevel models and provide biased estimates. To cope with this problem, Bayesian estimation is often recommended as an alternative to frequentist estimation, such as least squares or maximum likelihood estimation. This paper proposes a multilevel framework under a Bayesian approach to model municipal waste generation with hierarchical data structures. Using a real-world dataset of municipal waste generation in Denmark, the predictive accuracy of multilevel models is compared to aggregated and disaggregated Bayesian models using socio-economic external variables. Results show that Bayesian multilevel models outperform the other models in prediction accuracy, based on the leave-one-out information criterion. A comparison of the Bayesian approach with its frequentist alternative shows that the Bayesian model is more conservative in coefficient estimation, with estimates shrinking to the grand mean and broader credible intervals, in contrast with narrower confidence intervals produced by the frequentist models.

**Keywords**: Municipal waste generation; multilevel models; Bayesian data analysis; prediction

## 3.1 Introduction

Effective prediction of waste generation rates is an essential part of the design, implementation, and improvement of waste management operations (Ramos et al., 2018). From

a practical point of view, inaccurate predictions of the amount of generated waste can result in inefficient decisions regarding infrastructure, equipment capacity, or collection schemes. These decisions have a direct impact on the quality of service and the operating costs of a waste management system (Zbib and Wøhlk, 2019). The accuracy and relevance of predictions rely especially on the quality and structure of the underlying data used. In the context of municipal waste, data usually present structures comprising repeated observations over time from different municipalities or districts. Such data structures are usually referred as hierarchical or multilevel structures (Heck and Thomas, 2020).

A large variety of studies apply various methods to predict waste generation rates, and they differ in both methodology and sources of data. The methods used include descriptive statistics, regression analysis, material flow models, time series analysis, and artificial intelligence (Abbasi and El Hanandeh, 2016). These methods typically use either aggregated data from municipal or state entities or self-reported data collected by questionnaire (Hannan et al., 2015).

The majority of existing studies use regression analysis and model waste generation rates at the municipal or county level, using socio-economic and other external variables as predictors of future behavior (Abdoli et al., 2011). The main reasons this approach is widely used are because of the availability of data and the simplicity of this type of model. Usually, the data obtained from each municipality are analyzed by either pooling all municipalities into a single model or analyzing them separately. The main limitations of such approaches are their underlying assumptions. Including all municipalities in a single model assumes that all have similar behavior, with the risk of underfitting the data, whereas using individual models assumes that municipalities have nothing in common, with the risk of overfitting the data. In other fields, this problem is increasingly being dealt with by using multilevel models (Heck and Thomas, 2020). However, to the best of our knowledge, no previous study has applied a multilevel approach in the context of municipal waste generation.

Multilevel modeling is an extension of regression models and makes it possible to model data that have clustered or hierarchical structures. The main advantage of using a multilevel model is that it allows information to be pooled across clusters—municipalities in the present case—to improve the estimates of the parameters of the model (McElreath, 2020). This pooling means that each municipality helps to improve the estimates of the other municipalities and of the overall population. In general, traditional models can have one of two pooling structures: aggregated pooling or disaggregated pooling. An aggregated, or complete, pooling model assumes that there is no variability among municipalities and therefore it fits a single model shared by all municipalities (e.g., a simple linear regression). In contrast, the disaggregated pooling approach assumes that

municipalities do not share any relevant characteristics with each other, and therefore fits each one as a separate model. The selection of a model approach depends, like most decisions in data analysis, on the underfitting/overfitting trade-off (Gelman, 2006).

In contrast to the two traditional pooling approaches (aggregated and disaggregated pooling), a multilevel model uses a partial pooling strategy, which captures the systematic differences between municipalities by partitioning the variance into the *between-municipality* variance and the *within-municipality* variance. This allows each municipality to have a different average outcome, but the overall population average is also estimated by the model. This results in less underfitting than the complete pooling approach and less overfitting than the no-pooling approach; therefore the model produces better estimates (McElreath, 2020). However, increasing the complexity of the model can result in inaccurate estimates, particularly in the case of small sample sizes for each municipality (Gelman et al., 2013).

Bayesian data analysis has been increasingly used in a variety of fields in recent years, and has been recommended in preference to the traditional frequentist approach, particularly in the context of small sample sizes (Smid et al., 2020). Unlike the frequentist approach, Bayesian estimation is not based on the asymptotic behavior of the data, and thus results can be interpreted and validated for any sample size (Kaplan, 2014). The Bayesian approach has several advantages: First, it allows the integration of *a priori* knowledge using prior distributions in the parameters of the model. These prior distributions are then conditioned on the data, which is especially useful when expert knowledge is available. Second, the flexibility of a Bayesian model can be used to explicitly quantify the modeling uncertainty of the outcome. This flexibility can account for reduced sample sizes and can also include complex structures such as multilevel modeling (Miočević et al., 2017).

In this paper, we propose a multilevel framework using a Bayesian approach to predict municipal waste generation rates. We show that the proposed method has two main advantages over traditional modeling approaches. First, a multilevel framework allows us to obtain better estimations by incorporating the hierarchical structure of the data into the model. This is done by allowing correlation of observations in the same location over time, partitioning the variation into between-municipalities and within-municipalities components. Second, in a Bayesian approach, uncertainty can be modeled explicitly by the use of prior distributions for the model estimates, which produces more intuitive results.

Our approach is illustrated using a real-world dataset of annual waste generation rates from the 98 municipalities of Denmark for the period from 2010 to 2017. Several explanatory variables are tested to explain the variation between and within municipalities, including socio-economic and demographic variables. In addition, six waste types are

considered as response variables: general waste, burnable waste, glass, metal, cardboard, and plastic. General waste refers to mixed domestic waste collected from households. We separately determine the explanatory variables that influence each waste type. Then, we compare two multilevel models, varying intercept and varying slope, with the traditional aggregated and disaggregated models.

The remainder of this paper is structured as follows. Section 3.2 reviews the related literature. Section 3.3 presents the proposed methodology and provides details of each of the stages of the study, including the proposed Bayesian model. In Section 3.4, the results of our case study, using the proposed methodology, are shown. Finally, Section 3.5 concludes the paper.

## 3.2 Literature review

Municipal solid waste generation has been extensively studied in the literature, and a large variety of methods have been applied depending on the scale (e.g., household, municipal, state) and time period (short, medium, or long term) of the research. The most common approach used in studies of waste generation at macro-levels, such as the municipality, district, or country level, has been multiple regression analysis with utility maximization models.

Johnstone and Labonne (2004) use a panel dataset to analyze the determinants of solid waste generation using municipal solid waste, demographic, and economic data at a country level for 30 OECD countries from 1980 to 2000. Similarly, Callan and Thomas (2006) use a utility model to examine the demand for disposal and recycling services based on data for 351 municipalities in Massachusetts. In a multiple regression analysis framework, Hage et al. (2009) analyze the factors determining the generation rates of household plastic packaging for 252 municipalities in Sweden, focusing mainly on garbage pricing, socio-economic and demographic factors, and environmental preferences. Sidique et al. (2010) analyze the effects of various recycling and waste management policy variables on the recycling rate by utilizing municipality-level data from 86 municipalities in Minnesota from 1996 to 2004. The study uses a utility maximization model and accounts for the cumulative effect of the expenditure variable on the recycling rate. Lebersorger and Beigl (2011) use a multiple regression model to study waste generation rates based on socio-economic factors, including municipal tax revenue per capita, household size, and the percentage of buildings with solid fuel heating systems, from 542 municipalities in the Province of Styria, Austria. In the same fashion, Wei et al. (2013) use multiple linear regression to analyze waste generation at the national level in China. Oribe-Garcia et al. (2015) study socio-economic features relevant to waste generation for 112 municipalities

in Biscay, Spain, using a range of factor models. In that study, the authors account for differences in municipalities' waste generation by proposing two separate models, one for the overall region and a second with clustering of municipalities.

More recent studies have focused on the advantages of using artificial intelligence instead of traditional regression approaches to estimate waste generation rates. Azadi and Karimi-Jashni (2016) compare the performance of artificial neural networks (ANNs) and multiple linear regression to predict seasonal municipal waste generation rates. The accuracy of the two methods is compared based on a case study of 20 cities in the province of Fars, Iran. Similarly, Perera and Fernando (2020) compare the performance of ANNs and regression analysis using data from 15 local authorities from the districts of Colombo and Gampaha in Sri Lanka. Finally, Araiza-Aguilar et al. (2020) use multiple linear regression to study the effects of different social and demographic variables in 124 municipalities in Chiapas, Mexico. None of the above studies have explicitly considered the interactions between municipalities using a multilevel approach.

Bayesian data analysis has been successfully applied in a diverse range of fields, including behavioral sciences such as education, physiology, economics, and medicine (Kruschke and Liddell, 2018). However, in the area of waste management, only a few studies have considered the Bayesian approach; its application has been limited to survey data and it has not been applied in the context of macro-level analysis. Chu et al. (2016) use a Bayesian belief network model to determine the factors that affect the separation of waste for collection in China, including political, economic, social, cultural, and technological factors. Hoang et al. (2017) apply a Bayesian model average method combined with a multivariate linear regression to identify factors influencing household waste generation in Vietnam. Finally, Ceylan (2020) proposes a Bayesian Gaussian process regression model tuned by Bayesian optimization to forecast municipal solid waste generation in Turkey.

## 3.3 Methodology

This section presents the methodology used to study waste generation rates. The methodology is divided into two parts. First, Section 3.3.1 introduces the concept of frequentist data analysis. Second, Section 3.3.2 explains the basics of Bayesian data analysis, and the main differences to the frequentist approach. Next, Section 3.3.3 explains variable selection under a Bayesian approach and Section 3.3.4 specifies the approach used to impute missing values. Finally, Section 3.3.5 presents the metrics used to compare the prediction accuracy of the different models and their differences to frequentist approaches. Figure 3.1 summarizes the methodology used in this paper, divided into two parts. The model approach refers to the underlying approach, either frequentist or Bayesian. For each of

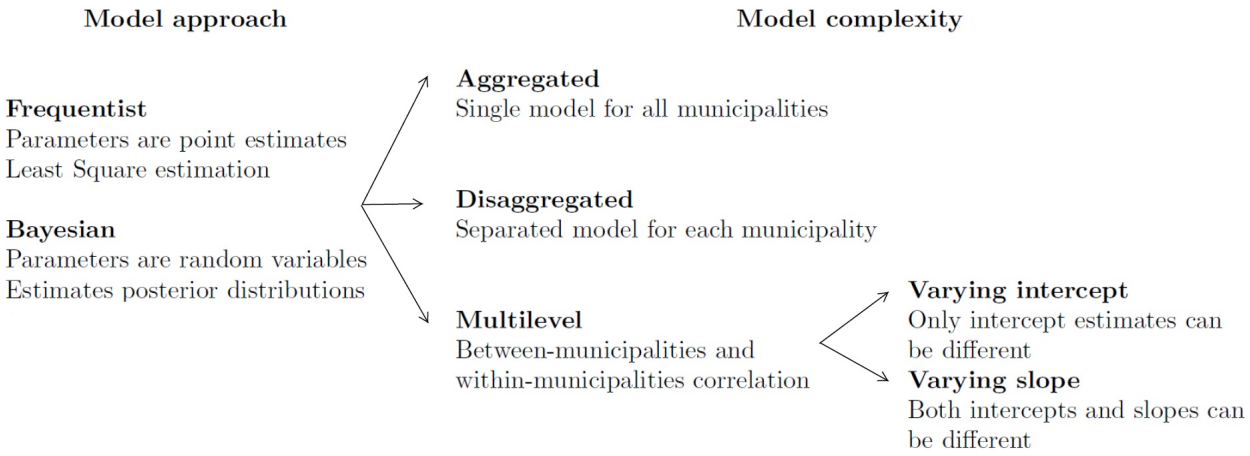


Figure 3.1: The model selection can be divided into modeling approach (frequentist or Bayesian) and model complexity (aggregated, disaggregated, or multilevel).

these, the model complexity corresponds to the assumptions used in the linear model: aggregated, disaggregated, or multilevel.

### 3.3.1 Frequentist data analysis

Frequentist data analysis is based on frequentist inference, using the relative frequency or proportion of events that occurs in a repeated experiment. This can be considered the "classical" approach to data analysis, which uses hypothesis testing, significance tests (p-values), and confidence intervals to perform statistical inference.

#### Frequentist linear models

The most commonly used method under the frequentist approach to predict waste generation rates using external variables is linear regression. Linear regression assumes that the $i$th observation, $i \in \{1, ..., N\}$, of a dependent variable $y_i$ has a linear relationship with an independent variable $X_i$:

$$y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{3.1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma)$ is the residual error. The model parameters in this case are considered point estimates (single, fixed values) and their estimation is performed by least squares estimation. The extension to a multiple linear model, for example, including the effects of age and immigration in the same model, is straightforward because it consists of including an extra parameter with its corresponding prior distribution for each external variable.

When the data are structured hierarchically, as in the case of municipal waste, one of three assumptions can be made. First, an *aggregated model* assumes that there is

no systematic difference between municipalities and, thus, includes all observations in a single linear model. Second, a *disaggregated model* assumes that municipalities do not share relevant characteristics and specifies a separate model for each municipality. Finally, a *multilevel model* explicitly incorporates the hierarchical structure of the data into the model.

**Frequentist multilevel model**

The multilevel model extends the aggregated and disaggregated models by allowing the parameters of the model to vary depending on the municipality the observation comes from. Part of the multilevel modeling process is to decide which parameters are considered to vary among municipalities and which are considered to be constant. In a linear model, the intercept and the slope are the parameters to be estimated for each independent variable, and these define the two variations of the multilevel model: the *varying intercept* and the *varying slope* models. Assuming there are $M$ municipalities, the *varying intercept* model can be described as:

$$y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \epsilon_{ij} \tag{3.2}$$

$$\beta_{0j} = \mu_0 + u_{0j} \tag{3.3}$$

$$\beta_1 = \mu_1 \tag{3.4}$$

where $j$ is the $j$th municipality, $j \in \{1, ..., M\}$. $\epsilon_{ij} \sim \mathcal{N}(0, \sigma)$ and $u_{0j} \sim \mathcal{N}(0, \sigma_0)$, usually referred to as *random effects*, are the residual errors. The multilevel model separates the between-municipalities variability ($\sigma$) and the within-municipalities variability ($\sigma_0$). In the *varying slope* model, the term $\beta_1$ in Eqs. (3.2) and (3.4) is replaced by:

$$\beta_{1j} = \mu_1 + u_{1j} \tag{3.5}$$

where $u_{1j} \sim \mathcal{N}(0, \sigma_1)$.

### 3.3.2 Bayesian data analysis

The Bayesian approach to data modeling is based on Bayesian inference, where the main characteristic is that each parameter of a model is a random variable (Gelman et al., 2013). This feature allows Bayesian models to explicitly model the underlying uncertainty of the estimation of a given parameter. Under this framework, the Bayes' theorem is used to

model the probability of a parameter $\theta$ given a data set $y$ as

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}. \tag{3.6}$$

Using this approach, we can estimate the probability distribution of a parameter, $p(\theta \mid y)$, which represents the relative plausibility of different values of the parameter, conditional on the data and the model (McElreath, 2020). The main result of the Bayesian analysis is that the probability distribution of a parameter, $p(\theta \mid y)$, referred to as the posterior distribution, is proportional to the product of the information contained in the data (likelihood), $p(y \mid \theta)$, and the information available before observing the data (prior), $p(\theta)$. The posterior distribution contains all the information needed to perform the Bayesian inference. Using this approach, Bayesian modeling requires the specification of a likelihood function for the data (e.g., $y_i \sim \mathcal{N}(\mu, \sigma)$) and a prior distribution for the parameters in the model (e.g., $\mu \sim \mathcal{N}(0, 1)$), followed by an estimation of the posterior distribution, usually using numerical techniques (Nalborczyk et al., 2019). The numerical techniques to fit the models are usually based on Markov Chain Monte Carlo (MCMC) simulations, for which many methods use a Gibbs sampler approach or a Hamiltonian sampler (Scott and Berger, 2010). For high-dimensional models such as multilevel models, Hamiltonian Monte Carlo is usually superior, and this is the sample technique used in this study.

### Bayesian linear models

In the context of municipal waste management, we are interested in modeling waste generation rates, which are usually measured in kilograms per person per unit of time. Thus, the target variable $y$ is a continuous variable with non-negative values, measuring the amount of waste. An exponential distribution is a proper selection for the likelihood distribution in this scenario, and this also simplifies the analysis by using a single parameter for the estimation. In the following models, the likelihood distribution for the waste variable $y$ is:

$$y_i \sim \text{Exponential}(\lambda), \tag{3.7}$$

where $\lambda$ is the parameter to be estimated using a link function to the external variables available in the dataset (socio-economic, demographic, or others). In the *aggregated* linear model, we obtain a single intercept and single slope for all municipalities, and the model can be formulated as follows:

$$y_i \sim \text{Exponential}(\lambda_i) \tag{3.8}$$

$$\log(\lambda_i) = \beta_0 + \beta_1 X_i \tag{3.9}$$

$$\beta_0 \sim \mathcal{N}(\mu_0, \sigma_0) \tag{3.10}$$

$$\beta_1 \sim \mathcal{N}(\mu_1, \sigma_1) \tag{3.11}$$

where $\lambda$ is modeled as a linear function of a continuous variable $X$, and the intercept $\beta_0$ and slope $\beta_1$ follow a—prior—normal distribution with the parameters $\{\mu_0, \sigma_0\}$ and $\{\mu_1, \sigma_1\}$, respectively. In general, to define a proper prior distribution of the parameters of the model ($\beta_0$ and $\beta_1$), we can include *a priori* information about the process (expert knowledge, for example), specify a weakly informative prior with a rather large variance, or use prior predictive simulations to make sure that the model predictions prior to seeing the data lie within the plausible outcome space (Stegmueller, 2013). In this study, we tune our priors using prior predictive simulations before updating the parameters using the data. Setting the prior of $\beta_0$ to $\mathcal{N}(3, 0.5)$ and $\beta_1$ to $\mathcal{N}(0, 0.3)$ makes the models treat extreme values with skepticism in the presence of scarce data.

In a *disaggregated* linear model, there is no pooling of information between municipalities, and it is assumed that each municipality can be modeled independently. In this approach, an individual model is fitted using Eqs. (3.8)–(3.11) for each municipality.

### Bayesian multilevel modeling

The most straightforward Bayesian multilevel model is the *varying intercept* model, in which we allow the intercept for each municipality, $j = 1...N$, to be different ($\beta_{0j}$), but preserve the same slope ($\beta_1$). The model can be defined as follows:

$$y_{ij} \sim \text{Exponential}(\lambda_{ij}) \tag{3.12}$$

$$\log(\lambda_{ij}) = \beta_{0j} + \beta_1 X_{ij} \tag{3.13}$$

$$\beta_{0j} \sim \mathcal{N}(\mu_0, \sigma_0) \tag{3.14}$$

$$\beta_1 \sim \mathcal{N}(\mu_1, \sigma_1) \tag{3.15}$$

$$\mu_0 \sim \mathcal{N}(3, 0.5) \tag{3.16}$$

$$\sigma_0 \sim \text{Exponential}(3) \tag{3.17}$$

where $\beta_{0j}$ follows a normal distribution with parameters $\mu_0$ and $\sigma_0$. The hyperparameter $\mu_0$ represents the average intercept among municipalities, $\sigma_0$ is the variability of the intercepts, and both hyperparameters are assigned their own distributions. We assign a $\mathcal{N}(3, 0.5)$ to $\mu_0$ and an Exponential(3) to $\sigma_0$. Finally, Eq. (3.12) defines the likelihood of the data, and Eq. (3.14) is the prior distribution of the parameter that describes the population intercepts. This varying intercept model is relevant when municipalities present a similar effect of a variable on the amount of waste, but they differ in the average impact.

We can extend the varying intercept model to a *varying slope* model by allowing the slope $\beta_1$ to be different for each municipality $j$. This is achieved by replacing Eq. (3.15) by:

$$\beta_{1j} \sim \mathcal{N}(\mu_1, \sigma_1) \tag{3.18}$$

and adding the hyperparameter distributions to the model:

$$\mu_1 \sim \mathcal{N}(3, 0.5) \tag{3.19}$$

$$\sigma_1 \sim \text{Exponential}(3) \tag{3.20}$$

A varying slope model can be useful in cases where, for example, some municipalities show an increase in waste generation with age, and some show a decrease. More generalized models include specification of correlation matrices' priors or include more than one level of hierarchy.

The selection of the adequate model complexity level depends on the amount and type of data, the overfit/underfit trade-off, and the objective of the study. An initial assessment to choose between a complete pooling, no pooling, or multilevel approach, can be based on the intraclass correlation coefficient (ICC). The ICC computes the proportion of the response variable's variance that is due to between-level differences (Mulder and Fox, 2019). In our case, the ICC represents the amount of variance from the waste type that is caused by differences between municipalities. Another index is the design effect index (Deff), which measures the inflation in variability of the estimates due to clustering, and it is often used as a rule of thumb to indicate whether multilevel structures should be used (Lai and Kwok, 2015). It is defined as $\text{Deff} = 1 + (n-1)\text{ICC}$ where $n$ is the average number of observations per cluster. Values of this index above 2 are usually considered

to be appropriate to model as multilevel.

### 3.3.3 Bayesian variable selection

Several variables have been used to explain variations in municipal solid waste. They mainly include demographic, weather, and socio-economic variables such as income, education, or gender (Abbasi and El Hanandeh, 2016). In a traditional approach, variable selection is a search problem in which the objective is to find a single optimal model from all the possible combinations of original variables by minimizing or maximizing a certain criterion. In contrast, the Bayesian approach is probabilistic, based on determining the probability that a variable should or should not be included in the model. This probability, is referred to as the posterior inclusion probability (George and McCulloch, 1997) and can be estimated as follows.

Given a response variable $y_i$, and $p$ explanatory variables with values $x_{i,j}$, $j = 1, ..., p$, the response variable can be modeled as the linear combination of $p$ variables with parameters $\theta_j$ for each variable:

$$y_i = \beta_0 + \sum_{j=1}^{p} \theta_j x_{i,j} + \epsilon_i, \tag{3.21}$$

where $\beta_0$ is the intercept, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the error term. The variable selection problem is to determine which regression parameters $\theta_j$ should be set to zero. To do this, an auxiliary indicator variable $I_j$ can be defined, with $I_j = 1$ if the explanatory variable $j$ is present in the model, and $I_j = 0$ if it is not. A second auxiliary variable is used to represent the effect size of the explanatory variable when $I_j = 1$, which is denoted by $\beta_j$, and therefore $\theta_j = I_j \beta_j$. The variable selection part of the model estimates $I_j$ and $\theta_j$. The variable $\beta_j$ can be defined in several ways, all of which result in different fitting methods (O'Hara and Sillanpää, 2009). After the model has been defined, the posterior inclusion probability for the variable $j$ can be computed as the average value of $I_j$, and it is usually fitted using a MCMC approach.

### 3.3.4 Imputation of missing data

Incomplete data is a common problem in most applications, and can limit the implementation and analysis of statistical and machine learning models (Lin and Tsai, 2020). The problem of missing data is especially relevant in the context of waste management, in which the lack of data can be due to several practical issues, such as errors in the measurements, system failures, or lack of reporting. There are two main approaches to deal with missing values, namely deletion and imputation (Garciarena and Santana, 2017).

Deletion methods ignore cases or variables in which there are missing values, and due to their simplicity these methods can be useful in cases with low rates of missing values (Lan et al., 2020). However, when the rate of missing values is high, deletion can cause a major loss of information and may cause bias and overfitting in the resulting models (Purwar and Singh, 2015).

Imputation for missing data with multilevel structures is usually performed before the modeling stage, using either a joint modeling approach or a fully conditional specification model (Grund et al., 2018). In the joint modeling approach, a single model is specified for all variables with missing data, whereas in the fully conditional specification model, the missing data are imputed separately for each variable (Carpenter and Kenward, 2012). In our application, because the explanatory variables are complete, we use a fully conditional specification model, which basically iterates univariate multilevel imputation of the variables. Specifications of the model and its implementation in the R software environment can be found in van Buuren and Groothuis-Oudshoorn (2010).

### 3.3.5 Model evaluation

After the Bayesian model has been fitted, its predictive accuracy is usually measured using cross-validation and information criteria approaches. The evaluation metrics can be used to select a single—best—model for the given data or to improve estimations by averaging different models, assigning weights to their posterior probabilities (Congdon, 2007). The most common methods for model comparison are the Bayesian information criterion (BIC), the deviance information criterion (DIC), the Akaike information criterion (AIC), and the leave-one-out information criterion (LOOIC) (Vehtari et al., 2017). In this study, we focus on the prediction accuracy of the fitted models, thus using on the LOOIC method. LOOIC estimates pointwise out-of-sample prediction accuracy using the log-likelihood evaluated at the posterior simulations of the parameter values. The Bayesian LOOIC predictive fit estimate is:

$$\text{LOOIC} = \sum_{i=1}^{n} p(y_i \mid y_{-i}), \tag{3.22}$$

where $p(y_i \mid y_{-i})$ is the leave-one-out predictive density obtained by fitting the data without the $i$th data point. Lower LOOIC values denote better out of sample predictive accuracy performance.

## 3.4 Case study

In this section, the proposed methodology is applied to a real-world dataset from Denmark. Section 3.4.1 describes the data set and summarizes its main statistics and the structure of its missing values. Section 3.4.2 presents the results of the model selection process, measuring predictive accuracy in terms of the LOOIC of the multilevel, aggregated, and disaggregated Bayesian models. Finally, Section 3.4.3 presents a comparison of the coefficient estimation of the multilevel Bayesian model with the alternative frequentist approach. The analysis was conducted using R 3.6.3 software and run on a 3 GHz Intel X5450 processor with 24 GB RAM. The multilevel frequentist models were run using the **lme4** R package (Bates et al., 2012) and the Bayesian analysis was conducted using the **brms** R package (Bürkner, 2017).

### 3.4.1 Data description

The dataset consists of yearly observations of municipal waste generation from all of the 98 municipalities in Denmark between 2010 and 2017. We use six waste types, which are: general waste, burnable waste, glass, metal, cardboard, and plastic. The data are reported by each municipality, and were provided for this study by the Ministry of the Environment and Food of Denmark. The dataset was combined with socio-economic variables obtained from Statistics Denmark (Statistics Denmark, 2019). The external variables include average taxable income, average age, gender (% of men), marital status (% of divorced individuals), immigration (% of immigrants), and educational attainment (% of individuals that graduated from a bachelors program). The last three variables were reported as the percentage of the total population of the municipality in the observation year. The selection of the external variables is based on both data availability and the previous studies that have found socio-economic variables to be relevant in the prediction of waste generation (Vu et al., 2019b; Kannangara et al., 2018; Kumar and Samadder, 2017). Other variables were also tested, including the number of farmhouses and number of households, but were not found to have a statistically significant correlation with waste generation. For details on the socio-economic variables used in this study see .

Municipalities in Denmark have high variability in the reported values of variables such as population, area, density, type of households, and waste generation. The average annual waste generation, as the total of the six waste types, was 331 kg per person, and most of this weight is general waste and burnable waste. Figure 3.2 shows the average waste generation by type for each of the 98 municipalities. As can be seen, there is large variation in the total waste amount. The proportion represented by each

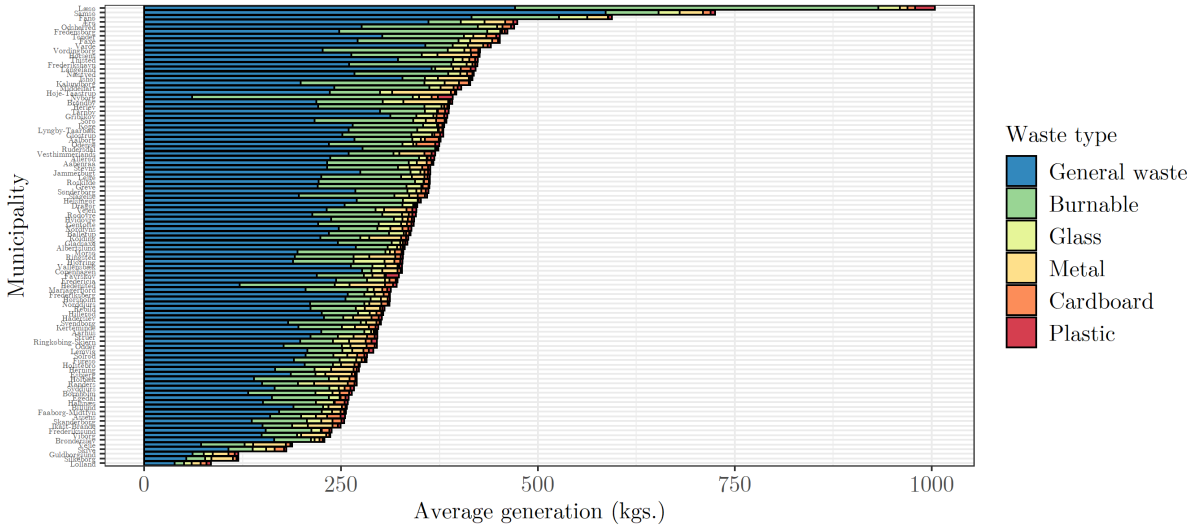


Figure 3.2: Average waste generation (kgs.) per person from 2010 to 2017 by municipality and by waste fraction.

waste type also varies among municipalities. Most notably, the proportions of general waste and burnable waste show much higher variation than the other waste types. For example, whereas general waste is the largest waste type in most municipalities, in some municipalities (such as Nyborg or Hedensted) burnable waste accounts for more than half of the average waste generation.

The minimum and maximum waste generation rates also differ immensely, from 113 kg per person to 1,054 kg per person. This difference may be due to the influence of small municipalities that are mainly used for summer residences, compared to municipalities that primarily comprise permanent residences, among which variation was lower. In terms of the correlation between the weights of each waste type, we generally found little correlation, with an average correlation coefficient of 0.2.

Whereas the dataset of socio-economic variables was complete for the period studied, some waste variables had high amounts of missing data; 30.14% of a total of 784 observations featured missing values. These missing values exhibit a rather random behavior, concentrated on neither any specific municipality nor any particular year. Figure 3.3 illustrates the pattern of missing values in the original dataset for two waste fractions: general waste, which has only 5% missing data, and glass, which has over 50% missing data. Each row represents a year, the columns represent the 98 municipalities, the gray boxes represent observed values, and white boxes represent missing values. In order to perform the multilevel modeling, the imputation method described in Section 3.3.4 was used.

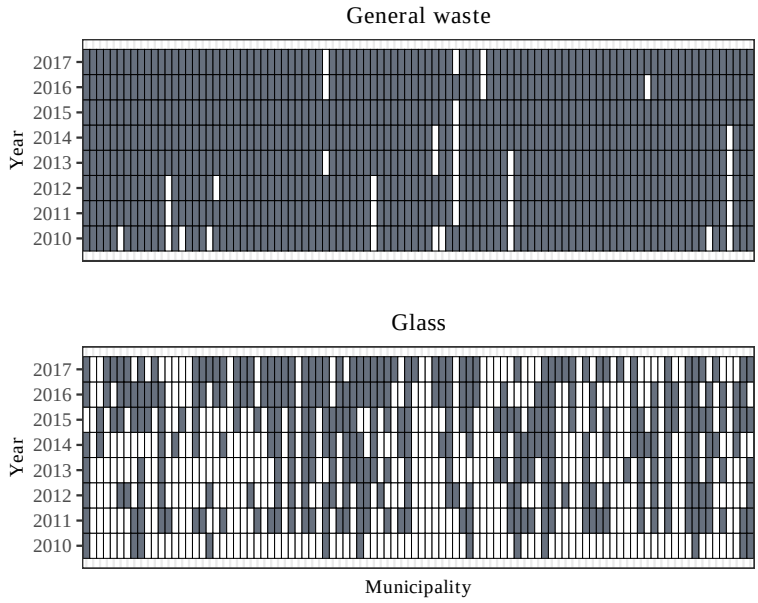


Figure 3.3: Patterns of missing values for general waste and glass, from the original dataset. Gray and white represent observed and missing values, respectively.

| Waste fraction | ICC | Deff | Estimate | SE |
|---|---|---|---|---|
| General waste | 0.51 | 4.60 | 224.84 | 8.04 |
| Burnable | 0.44 | 4.05 | 77.13 | 5.79 |
| Metal | 0.29 | 3.03 | 15.28 | 1.35 |
| Plastic | 0.29 | 3.02 | 3.66 | 0.33 |
| Glass | 0.23 | 2.60 | 15.96 | 0.68 |
| Cardboard | 0.23 | 2.60 | 6.91 | 0.38 |

Table 3.1: All variables present a design effect index (Deff) above 2. General waste and burnable show high intraclass correlations.

### 3.4.2 Model selection

Three different model types were compared for each waste type: aggregated, disaggregated, and multilevel modeling. In addition, two types of multilevel approach were tested: the varying intercept and varying slope models. These approaches were tested using different combinations of external variables, based on their inclusion probabilities. Before defining the appropriate models for each waste type, we tested the ICC and Deff to evaluate the differences in responses from each municipality. In Table 3.1, we show the results for each waste type, including the estimated mean response and standard error (SE). Our results show that all variables have high variability between municipalities (Deff value above 2), indicating that a multilevel approach is a suitable choice for all variables, and especially so for the variables for general waste and burnable waste.

| Waste type | Explanatory variables | Best model |
|---|---|---|
| General waste | Age | ML Varying slope |
| Burnable | Age | ML Varying slope |
| Metal | Immigration | ML Varying slope |
| Plastic | Age, gender, education | ML Varying intercept |
| Glass | Age, immigration | ML Varying intercept |
| Cardboard | Gender, marital status, education, immigration | ML Varying intercept |

Table 3.2: Best model for each waste type, including explanatory variables and type of model. ML: Multilevel.

Six external variables were considered as candidates for model selection: age, immigration, gender, education, marital status, and income. For each of the waste types, different combinations of external variables were tested, considering the inclusion probabilities obtained as described in Section 3.3.3. For each combination of variables, the aggregated, disaggregated, varying slope, and varying intercept models were tested using both the frequentist and Bayesian approaches.

Table 3.2 summarizes the results of the model selection phase. For each waste type, the selected explanatory variables and the best model are reported. For all waste types, the multilevel approach was the best-performing in terms of LOOIC. For general waste, burnable, and metal, the varying slope model is the best model, which shows that the effect of the selected variable can be considered to vary between the different municipalities. For plastic, glass, and cardboard, the varying intercept model is best. In terms of the explanatory variables, age is relevant to most waste types (general waste, burnable, plastic, and glass) and immigration is relevant to three (metal, glass, and cardboard). Gender and education are only relevant to plastic and cardboard, and marital status is only relevant to cardboard. Taxable income was not found to be an explanatory variable for any of the waste types. In terms of variable transformations, we standardize the variable age by using its logarithm.  provides details on the model selection process for general waste.

A comparison of the prediction accuracy between the best multilevel model (varying intercept or varying slope) and the aggregated and disaggregated Bayesian models is presented in Table 3.3. For each waste type, Table 3.3 shows the LOOIC estimate, the standard error (SE) of the estimate and the standard error of the difference to the best model ($SE_d$). For all waste types, the best multilevel model is the model with the lowest LOOIC value. However, it is worth noting that for some waste types the accuracy of the three models are similar, as can be seen from the SE and the $SE_d$ values. This is the case for general waste, in which the SEs of the two models are SE = 26 and SE = 25,

| Waste type | Best multilevel | SE | Aggregated | SE | $SE_d$ | Disaggregated | SE | $SE_d$ |
|---|---|---|---|---|---|---|---|---|
| General waste | 10,037 | 26 | 10,053 | 25 | 36 | 10,850 | 27 | 37 |
| Burnable | 8,190 | 53 | 8,348 | 48 | 72 | 9,585 | 137 | 147 |
| Metal | 5,373 | 86 | 5,837 | 74 | 113 | 7,185 | 111 | 140 |
| Plastic | 3,438 | 69 | 3,484 | 83 | 108 | 4,803 | 204 | 215 |
| Glass | 5,883 | 41 | 5,884 | 42 | 59 | 6,247 | 55 | 69 |
| Cardboard | 4,427 | 62 | 4,530 | 57 | 84 | 5,211 | 110 | 126 |

Table 3.3: LOOIC for the best model and the corresponding model using the aggregated and disaggregated Bayesian models.

respectively, which are similar the standard error of the difference, $SE_d = 36$.

In the Bayesian approach, a full posterior distribution is obtained for each of the parameters of the model. In Table 3.4, the main statistics are shown for the posterior distributions of the varying slope model when predicting general waste based on age. Because the multilevel model allows each municipality to have different estimates of the parameters, results can be divided into municipality-level and population-level effects. At the municipality level, the standard deviations of the parameters represent the variance of the estimates between municipalities, compared to the overall average response at the population level. On average, the population-level effect of age on general waste is positive, with a unit change in kgs per 1.05 units of age. However, the standard deviation of the effect of age between municipalities is quite large (1.19), showing that the effect of age can be very different depending on the municipality. This result is obtained with high uncertainty in the standard deviation estimate, represented by an SE of 1.07, which can be explained by the small sample size of eight observations per municipality in this case study.  presents a histogram of the posterior samples of the intercept and the slope of the model.

### 3.4.3 Comparison of frequentist and Bayesian approaches

A major difference between frequentist and Bayesian approaches is the estimation of confidence intervals for the parameters of the model. In the frequentist approach, estimates are obtained based on the least squares or maximum likelihood methods, and then a confidence interval is constructed using the associated standard errors. In contrast, in the Bayesian approach a full posterior distribution of the parameters is obtained, and the confidence intervals—called credible intervals—are the quantiles of that distribution (usually 5% and 95% quantiles) (Stegmueller, 2013). Figure 3.4 compares the Bayesian and frequentist multilevel models predicting general waste based on age. The points represent the intercept for each municipality, surrounded by 95% credible/confidence intervals. The figure shows that for predicting municipality-level general waste, model choice is clearly

| Municipality-level effects: | | | | | |
|---|---|---|---|---|---|
| | | Estimate | SE | Lower-95% CI | Upper-95% CI |
| $\text{sd}(\beta_{0j})$ | Intercept | 1.06 | 1.04 | 1.00 | 1.17 |
| $\text{sd}(\beta_{1j})$ | Age | 1.19 | 1.07 | 1.03 | 1.36 |
| $\text{corr}(\beta_{0j}, \beta_{1j})$ | (Intercept, age) | 0.27 | 0.56 | -0.89 | 0.98 |
| Population-level effects: | | | | | |
| | | Estimate | SE | Lower-95% CI | Upper-95% CI |
| $\beta_{0j}$ | Intercept | 214.86 | 1.04 | 198.34 | 230.44 |
| $\beta_{1j}$ | Age | 1.05 | 1.05 | 0.95 | 1.15 |

Table 3.4: Estimates of the parameters of the varying slope model predicting *general waste* based on *age* drawn from their posterior distributions. sd(): standard deviation, corr(): correlation, CI: credible interval.

not trivial. From the bottom to the top of the plot, the frequentist model predictions range from around 110 to 325 kg, whereas the Bayesian model estimates are rather more consistent, at between 200 and 225 kg.

Compared to its frequentist equivalent, the conservatism of the Bayesian multilevel model results in rather extreme differences in predictions for some municipalities. For example, the frequentist maximum likelihood estimate for Nyborg is around 110 kg of general waste, whereas the corresponding mean posterior estimate is around 205 kg, almost twice as large. For the municipality Varde, the frequentist model estimate of the average is around 305 kg, whereas the multilevel model estimate is around 220 kg. The substantial differences for these municipalities arise because the Bayesian multilevel model is more skeptical about extreme predictions. It treats mean values of municipalities that are far from the overall mean as likely to be chance events. This treatment is based on a low estimated variation between the municipalities and the relatively small sample sizes in municipalities with extreme values of general waste production. To mitigate the risk of being misled by chance results in these municipalities, the Bayesian model aggressively shrinks municipality-level estimates of general waste towards the grand mean.

Finally, the two models yield different sizes of confidence/credibility intervals for estimated average general waste. The narrow confidence intervals of the frequentist multilevel model reflect a narrow range of values of average general waste that is consistent with an $\alpha = 5\%$ test. Conversely, the credible intervals of the Bayesian model encompass a wider range of values that are compatible with the model, the data, and the prior. For instance, for Nyborg municipality the model, data, and prior are consistent with a 95% credibility interval from 160 to 240 kg. Again, this reflects the more conservative nature of Bayesian estimation, where penalization through adaptive regularization helps protect

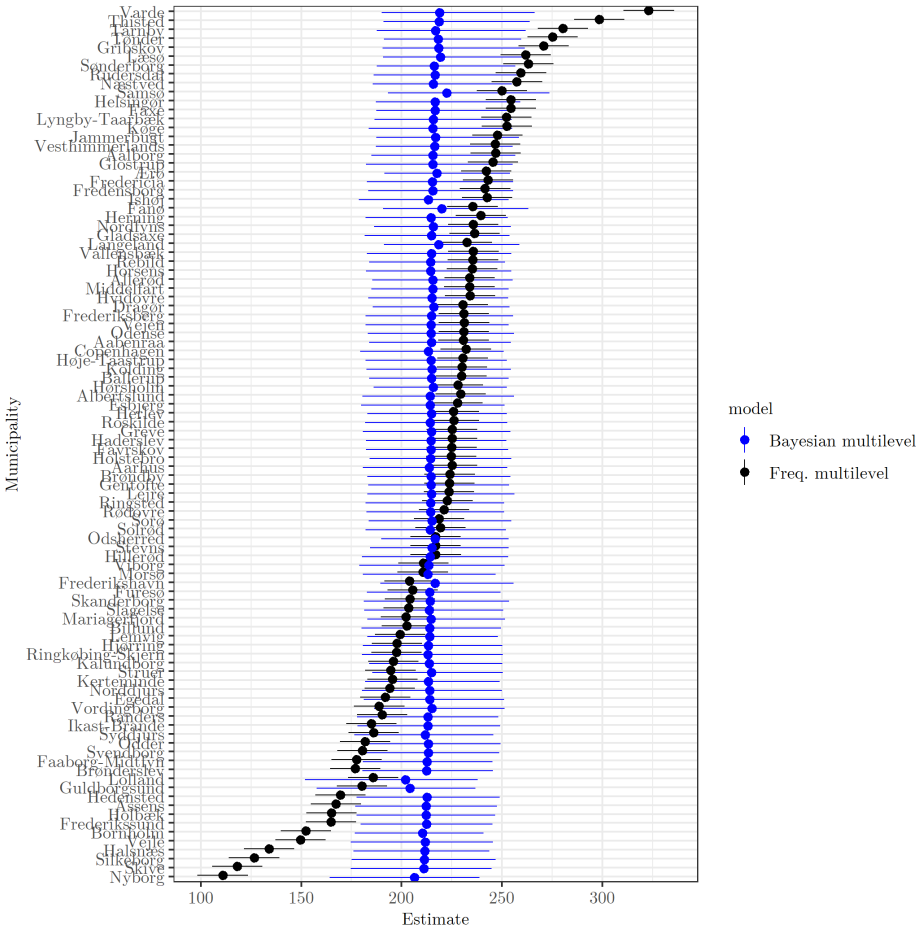


Figure 3.4: Comparison of the estimates and confidence/credible intervals of the intercept ($\beta_0$) for *general waste* based on *age* for the frequentist and Bayesian multilevel models. Results are reported for all 98 municipalities.

against overconfidence in the model results.

## 3.5 Conclusion

The present study proposes a multilevel framework under a Bayesian approach to predict waste generation rates in cases where data are structured hierarchically in municipalities or districts. This study is the first to investigate the advantages of using a multilevel Bayesian approach compared to aggregated and disaggregated linear models in the context of waste management. The proposed methodology is used in a case study of yearly waste generation rates for six waste types from the 98 municipalities of Denmark, combined with socio-economic and demographic variables in the period from 2010 to 2017.

In terms of prediction performance, the Bayesian multilevel model outperformed the traditional aggregated and disaggregated models for all waste types. For three of the six waste types analyzed, the varying intercept model performed best, and for the other three, the varying slope model performed best. The number of explanatory variables used in the best selected model also varies between the waste types, from a single external variable in the cases of general waste, burnable waste, and metal, to four variables in the case of cardboard. These results suggest that the selection of the model complexity has to be performed carefully because the optimal model may vary between different waste types.

Results and conclusions using Bayesian data analysis can be in conflict with those of the traditional frequentist approach. This has been the case in many previous studies (Nalborczyk et al., 2019). In the present study, we found that the Bayesian approach tends to treat extreme predictions with greater skepticism, which shrinks the estimates at the municipality level towards the grand mean. This more conservative behavior is also reflected by broader credible intervals, in contrast with narrower confidence intervals in the frequentist approach.

Some limitations of the proposed methodology are worth noting. First, the resulting posterior distributions of the parameters of the model depend heavily on the selection of the prior distributions. Selecting prior distributions that can translate expert knowledge into the model is essential in the modeling phase. Future research should investigate the impact of using, for instance, weakly informative priors compared to prior predictive simulations. Second, the implementation of multilevel Bayesian models can be computationally expensive for models with a large number of parameters. This can impose practical limitations particularly in the case when several models have to be compared.

Finally, our study provides a framework for future studies to assess the effects of different configurations of real-world waste datasets on prediction performance. For instance, the effects of the number of municipalities and the number of observations per

municipality should be studied further. The methodology presented in this paper can be extended to scenarios with two or more levels of hierarchy. A multilevel model with several levels can be used to study intra-class effects between, for example, countries and municipalities at the same time.


# Acknowledgements

This project was funded by the Danish Council for Independent Research - Social Sciences, project 'Transportation issues related to waste management' [grant number 4182-00021] and the AUFF NOVA grant for the project 'Opening the black box: Making Machine Learning lnterpretable for Organizational Research' [grant number AUFF-E-2019-9-3]. This support is gratefully acknowledged. We also thank the anonymous reviewers for their valuable comments.


# Appendix A

Table 3.5 provides an overview of the socio-economic variables used in the case study. Percentage variables represent the percentage of individuals over the total population of the municipality. Educational variables represent the highest educational level achieved.

| Socio-economic variable | Mean | St. Dev. | Min. | Max. |
|---|---|---|---|---|
| Population | 57,424 | 66,531 | 1,793 | 602,481 |
| Income (DKK) | 223,953 | 34,632 | 177,544 | 408,095 |
| Age | 42.22 | 2.61 | 35.90 | 52.80 |
| Divorced (%) | 8.66 | 1.31 | 5.72 | 12.04 |
| Primary education (%) | 21.23 | 3.67 | 10.58 | 29.92 |
| Secondary education (%) | 4.96 | 1.93 | 2.40 | 12.97 |
| Bachelor's education (%) | 0.81 | 0.77 | 0.22 | 5.36 |
| Immigrants (%) | 7.19 | 3.50 | 3.09 | 23.10 |
| Men (%) | 0.50 | 0.01 | 0.47 | 0.51 |
| Farmhouses | 11,789 | 7,416 | 785 | 39,471 |
| Households | 30,438 | 35,117 | 2,355 | 314,080 |

Table 3.5: Summary statistics of the municipal socio-economic variables.

# Appendix B

This appendix presents details of model selection for both the variables used and model complexity. First, Table 3.6 shows the inclusion probabilities for each of the six socio-

economic variables considered in the study for each waste type. Based on the variables showing the higher inclusion probabilities, Table 3.7 presents the LOOIC results of different model combinations for general waste. The best model is that with the lowest LOOIC, which is obtained by the varying slope model using only the variable age as an explanatory variable.

| Variable | General waste | Burnable | Metal | Plastic | Glass | Cardboard |
|---|---|---|---|---|---|---|
| Age | 1.00 | 1.00 | 0.21 | 1.00 | 0.97 | 0.28 |
| Marital status | 1.00 | 0.04 | 0.33 | 0.33 | 0.57 | 1.00 |
| Immigration | 1.00 | 0.03 | 1.00 | 0.73 | 0.89 | 1.00 |
| Education | 1.00 | 0.03 | 0.86 | 0.91 | 0.12 | 1.00 |
| Income | 0.47 | 0.03 | 0.39 | 0.80 | 0.12 | 0.30 |
| Gender | 0.31 | 0.08 | 0.99 | 1.00 | 0.11 | 1.00 |

Table 3.6: Inclusion probabilities of socio-economic variables for each waste type.

| Model | LOOIC | SE |
|---|---|---|
| $\log(\lambda_{ij}) = \beta_0 + \beta_1 * \text{age}_{ij}$ | 10,053 | 25 |
| $\log(\lambda_{ij}) = \beta_{0j} + \beta_1 * \text{age}_{ij}$ | 10,052 | 25 |
| $\mathbf{\log(\lambda_{ij}) = \beta_{0j} + \beta_{1j} * \text{age}_{ij}}$ | **10,037** | **26** |
| $\log(\lambda_{ij}) = \beta_0 + \beta_1 * \text{age}_{ij} + \beta_2 * \text{divorced}_{ij}$ | 10,053 | 25 |
| $\log(\lambda_{ij}) = \beta_0 + \beta_1 * \text{age}_{ij} + \beta_2 * \text{divorced}_{ij} + \beta_3 * \text{immigrant}_{ij}$ | 10,051 | 25 |
| $\log(\lambda_{ij}) = \beta_0 + \beta_1 * \text{age}_{ij} + \beta_2 * \text{divorced}_{ij} + \beta_3 * \text{immigrant}_{ij} + \beta_4 * \text{education}_{ij}$ | 10,051 | 25 |
| $\log(\lambda_{ij}) = \beta_0 + \beta_1 * \text{education}_{ij} + \beta_2 * \text{immigrant}_{ij} + \beta_3 * \text{divorced}_{ij}$ | 10,060 | 26 |

Table 3.7: LOOIC values for the best models, tested for general waste based on the inclusion probabilities.

# Appendix C

Figures 3.5 and 3.6 show histograms of the posterior samples of the intercept and slope, respectively, obtained for general waste modeled by age.

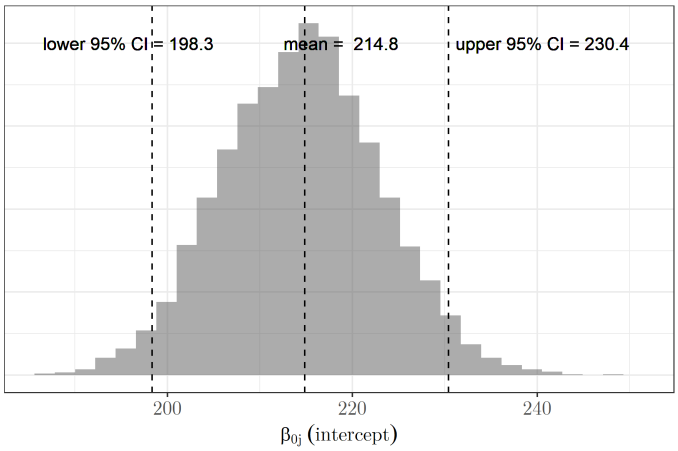


Figure 3.5: Histogram of the posterior samples for the intercept of the varying slopes model of general waste modeled by age.

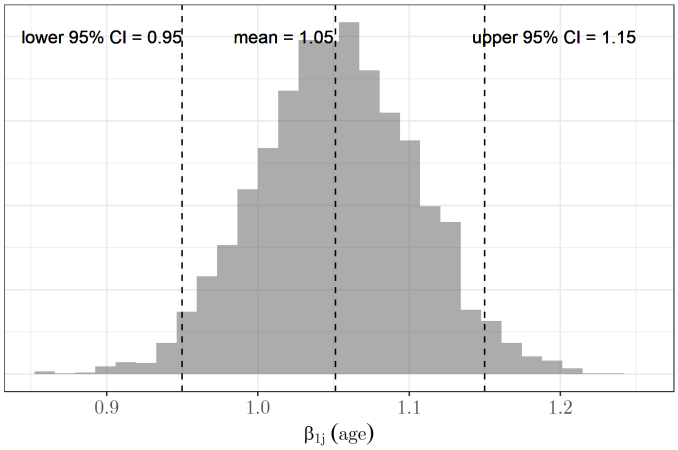


Figure 3.6: Histogram of the posterior samples for slope of the varying slopes model of general waste modeled by age.