Temporal Adjusted Prediction for Predicting Indian Reserve Populations in Canada

by

Derek Cho

B.Sc., The University of British Columbia, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

April 2018

© Derek Cho, 2018
Abstract

In order to predict the population of Indian reserves in Canada for the 2016 Census, we can construct a suitable model using data from the Indian Register and past censuses. Linear mixed effects models are a popular method for predicting values of responses on longitudinal data. However, linear mixed effects models require repeated measures in order to fit a model. Alternative methods such as linear regression only require data from a single time point in order to fit a model, but it does not directly account for within-individual correlation when predicting. Since we are predicting the responses of the same set of individuals, we can expect responses at the next time point to be strongly correlated with past responses for an individual.

We introduce a new method of prediction, temporal adjusted prediction (TAP), that addresses the issue of within-individual correlation in predictions and only requires data from a single time point to estimate model parameters. Predictions are based on the last recorded response of an individual and adjusted based on changes to the values of their covariates and estimated regression coefficients that relate the response and the covariates. Predictions are made using a random intercept model rather than a linear regression model. It is shown that if the random intercept accounts for a larger proportion of the random variation in the data than the random error term, then temporal adjusted prediction achieves a lower mean squared prediction error than linear regression.

TAP performs better than linear regression when predicting on the same set of individuals at different time points. It also shows similar prediction performance compared to linear mixed effects models estimated with maximum likelihood estimation despite only requiring data from one time point in order to fit a model.
Lay Summary

This thesis introduces a new method of prediction for populations of Indian reserves in Canada in 2016. Temporal adjusted prediction uses the population of a reserve from the previous census in 2011 as a starting point for prediction and extrapolates based on the changes in population from an auxiliary data source between the census dates in 2011 and 2016. In this case, the on-reserve populations of a band from the Indian Register are used as the auxiliary data source since it is highly correlated to the population of a reserve according to the census. One of the main benefits of temporal adjusted prediction is that it only requires data from a single time point to fit a model, while other similar methods require repeated measures.
Preface

This dissertation is based on unpublished work the author, Derek Cho, performed during a co-op work term at Statistics Canada. The temporal adjusted prediction method was developed by the author and used on data from the 2006, 2011 and 2016 Censuses and Indian Register at Statistics Canada under the supervision of Scott McLeish and Rose Evra. Further work on the theoretical properties in Chapter 2, and simulations in Chapter 3 were done at UBC under the supervision of Professor Gabriela Cohen-Freue.
# Table of Contents

Abstract ................................................................. ii  
Lay Summary ......................................................... iii  
Preface ................................................................. iv  
Table of Contents ................................................... v  
List of Tables ......................................................... vii  
List of Figures ........................................................ viii 
Acknowledgments ..................................................... x  

1 Introduction ......................................................... 1  
1.1 Related Work ................................................... 2  

2 Temporal Adjusted Prediction ................................. 5  
2.1 Notation .......................................................... 5  
2.2 Linear Regression ............................................. 6  
2.3 Linear Extrapolation ......................................... 9  
2.4 Temporal Adjusted Prediction ............................. 12  
2.4.1 Random Intercept ........................................ 15  
2.4.2 Comparing the MSPE of Linear Regression and Temporal Adjusted Prediction .................. 16  
2.4.3 Robust Temporal Adjusted Prediction ............... 20
3 Simulation Study .................................................. 23
  3.1 Data Generation ............................................. 24
  3.2 Simulation Results ........................................... 25
    3.2.1 Default Settings ..................................... 25
    3.2.2 Outliers ............................................. 26
    3.2.3 Additional Covariates ................................. 32
    3.2.4 Additional Time Point ................................. 35

4 Case Study: Prediction of Populations of Indians Reserves in Canada 41
  4.1 Background .................................................. 42
    4.1.1 The Indian Register .................................. 42
    4.1.2 Differences Between the Indian Register and Census Populations .......................... 43
  4.2 Data Description ........................................... 46
    4.2.1 Out-of-Scope Bands .................................. 47
    4.2.2 Multi-Band Reserves and Multi-Reserve Bands ....... 48
  4.3 Prediction of Indian Reserve Populations in 2016 .......... 51
    4.3.1 Exploratory Analysis ................................. 51
    4.3.2 Prediction Results ................................... 54

5 Conclusion ...................................................... 59

Bibliography ..................................................... 61
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Mean Squared Prediction Error of Temporal Adjusted Prediction and Linear Regression when varying Standard Deviation of Intercept</td>
<td>19</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Comparison of Median MSPE under Default Settings</td>
<td>25</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Comparison of Median MSPE with 5% Outliers</td>
<td>29</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Comparison of Median MSPE with 10% Outliers</td>
<td>31</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Comparison of Median MSPE with Additional Covariates</td>
<td>34</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>Comparison of Median MSPE with Data with an Additional Time Point under three Scenarios</td>
<td>37</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Levels of Census Metropolitan Influenced Zone</td>
<td>47</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>List of Multi-Band Reserves and Multi-Band Multi-Reserve Clusters</td>
<td>50</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Cross Tabulation of Province and Weighted Census MIZ</td>
<td>52</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Estimated Robust Regression Coefficients for 2006 and 2011 Reserve Populations</td>
<td>53</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Mean Square and Median Absolute Prediction Errors of 2016 Indian Reserve Populations</td>
<td>57</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Prediction of Selected Observation at Second Time Point via Linear Regression</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>Prediction of Selected Observation at Second Time Point via Extrapolation</td>
<td>10</td>
</tr>
<tr>
<td>2.3</td>
<td>Prediction of Selected Observation at Second Time Point via Temporal Adjusted Prediction</td>
<td>14</td>
</tr>
<tr>
<td>2.4</td>
<td>Mean Square Prediction Error of Temporal Adjusted Data Using Temporal Adjusted Prediction and Linear Regression</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>Potential Outliers in the Indian Reserves Population Data</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>Comparison of MSPE under Default Settings</td>
<td>26</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison of MSPE under Default Settings (Zoomed In)</td>
<td>27</td>
</tr>
<tr>
<td>3.3</td>
<td>Comparison of Mean Squared Prediction Error with 5% Outliers</td>
<td>28</td>
</tr>
<tr>
<td>3.4</td>
<td>Comparison of Mean Squared Prediction Error with 5% Outliers (Zoomed In)</td>
<td>29</td>
</tr>
<tr>
<td>3.5</td>
<td>Comparison of Mean Squared Prediction Error with 10% Outliers</td>
<td>31</td>
</tr>
<tr>
<td>3.6</td>
<td>Comparison of Mean Squared Prediction Error with 10% Outliers (Zoomed In)</td>
<td>32</td>
</tr>
<tr>
<td>3.7</td>
<td>Comparison of Mean Squared Prediction Error with an Additional Variable</td>
<td>33</td>
</tr>
<tr>
<td>3.8</td>
<td>Comparison of Mean Squared Prediction Error with an Additional Variable (Zoomed In)</td>
<td>34</td>
</tr>
<tr>
<td>3.9</td>
<td>Comparison of Mean Squared Prediction Error with Data with an Additional Time Point: Scenario (1)</td>
<td>37</td>
</tr>
</tbody>
</table>
Acknowledgments

There are several people that I would like to thank for helping me in my studies. In each step of my educational journey, I have relied on the help of others to get me through. Without their help, I would certainly not have come as far as I have. Firstly, I want to thank my family for their constant support of my educational endeavours. Without their love and encouragement, none of this would be possible.

Next, I would like to thank my supervisor, Professor Gabriela Cohen-Freue, for her guidance, patience and mentorship. From the time of my undergraduate studies, she has provided me several valuable opportunities for which I am truly grateful. I would also like to thank my second reader Professor Matias Salibian-Barrera.

Thanks to my past and current officemates Creagh, Jonathan, Nikolas, Saif, Sohrab and Yijun for your friendship. I will miss all of the conversations that we had and the trips to the village for dinner together. I also want to thank Danny, Qiong, Derek and the rest of my classmates for helping make grad school such an enjoyable experience. Thanks to Andrea, Ali and Peggy for all of your help and patience when answering my questions.

A special thanks to my colleagues at Statistics Canada where I did my co-op work term. I would like to thank Scott McLeish, Rose Evra and Tristan Cayn for being an awesome team of supervisors that I truly enjoyed working with. I will miss our end of the work day conversations. I also want to thank Susan Wallace, Sylvie Godin, Jarod Dobson and the rest of the Social and Aboriginal Statistics Division for their help with my projects while I was at Statistics Canada. Thanks to Eric McGregor from INAC for being helpful and providing me with the necessary data from the Indian Register for this thesis. I would also like to thank Pamela Best
for her kindness and willingness to help when I was in a tough spot.

Lastly, thanks to my friends Anthony, Gilbert, Louie, Raymond and Simon who have been with me since elementary school. I would also like to thank Lydia for her companionship and constant support of my studies. Thanks for coming to Ottawa to visit me while I was on co-op.
Chapter 1

Introduction

Linear regression models are a popular method for estimating the effects that explanatory variables have on a response. This relationship can then be used to predict future responses given new observations. In particular, a linear model can be fit to past data and then used to predict future responses given new values of the covariates. In many applications, if you are able to record new values of an explanatory variable, responses at the same time point will likely be recorded as well. However, this is not true in every case. For example, if you take data from two related datasets, one of which that has no recorded data for the current time point, you can predict the missing values from the other dataset using the estimated relationship between these datasets based on the past. One such application is the prediction of Indian reserve populations in Canada in 2016 using data from the Indian Register (IR) and the Census of Population. The goal is to use past data from the 2011 Indian Register and the 2011 Census to construct a model and then use the data from the 2016 Indian Register to make predictions of the populations of Indian reserves in Canada in 2016. Counts from census collection that are noticeably different from model predictions could warrant further investigation if the reason for the disparity isn’t immediately clear. This motivating example will be referred to throughout this thesis.

The Indian Register is an administrative database that contains information on all Registered Indians in Canada including band membership and whether or not they live on a reserve[5]. The population of most Indian reserves in Canada are
comprised mostly of Registered Indians[20] [16]. Thus, the number of individuals belonging to a First Nation or band, also referred to as band members, living on a reserve associated with that First Nation according to the IR should be approximately equal to the actual population of the reserve as measured by the census in most cases. The IR is useful for the prediction of reserve populations because it contains up to date information on the population changes of reserves since the previous census. The census and the IR populations of reserves for a given census year are highly correlated since a person listed as living on-reserve according to the register is likely to appear on the census as well. For most reserves, the census population is very similar to the IR on-reserve population at the time of each census, but for others, there are substantial differences between these two populations. Some reserves have people that are listed as living on-reserve according to the IR but do not appear on the census. Other reserves have people that appear on the census, but are not listed as band members in the IR. Statistical methods are necessary in order to model the relationship between the census and IR population and to produce accurate predictions.

In Chapter 2, we discuss two simple methods, linear regression and linear extrapolation which can be estimated using data from a single time point and used to predict future responses. Chapter 2 also introduces temporal adjusted prediction (TAP) and its robust variant which are able to predict using a model also fit with data from single time point while addressing the shortcomings of linear regression and linear extrapolation. Chapter 3 shows the results of a simulation study comparing temporal adjusted prediction to other models for predicting longitudinal data in various situations. In Chapter 4, we examine the case study in our motivating example by comparing the prediction accuracy of temporal adjusted prediction against that of selected methods for the 2016 Census population using data from the previous censuses and the IR. Lastly, Chapter 5 discusses conclusions and future directions for the work presented in this thesis.

1.1 Related Work

In the current literature, linear mixed effects (LME) models are a popular method for fitting data with repeated measures. LME models are suitable for longitudinal
data because they allow the modelling of both between and within-individual variation [10]. The effects of some variables may be common across individuals (fixed effects), while others may be specific to an individual (random effects). However, in order to estimate the random effects of each individual, multiple measurements or repeated measures for each individual must be available. If Indian reserve populations are measured at multiple time points, we can use the census and IR data to fit a LME model.

However, data from the IR is only available between 2006 and 2016. Due to the limited amount of time points for which complete data is available, we want to use a model that can be fit using complete data from only one time point. In this situation, we only have complete data for the 2006 and 2011 census years. That way, we can fit a model using 2006 Census and IR data to predict Indian reserve populations for 2011. Since we have the true populations for 2011 from the 2011 Census, we can use this as a validation set to test how well our model predicts. If we fit an LME model using both 2006 and 2011 data, then we are left without a validation set to test how well the model predicts since there is no response for 2016. Because of this, we are looking to use a method which can be fit with complete data from at least one time point rather than at least two.

Although a linear regression model estimated by least squares using data from a single time point can be used to predict future responses, it does not directly take into account the fact that the responses of specific individuals are correlated with their past responses. In the context of the Indian reserve population data, the individuals in the data are reserves and the covariates are the population of the reserves according to the IR in each census year and geographic variables such as the province in which the reserve is located. It makes sense that the population of a reserve in 2016 will be correlated with its population in 2011. This issue is further discussed in Section 2.2.

Another issue that this thesis aims to address is the effect that outliers have on parameter estimates. It is well known that even a small amount of outliers in the data can have significant effects on parameter estimates[21]. In the Indian reserve populations dataset, there are a number of observations which are considered outliers due to their distance from the bulk of the data. Thus, a robust procedure is needed to estimate parameters without being affected by contaminated
data. Several robust regression estimation methods exist such as M-estimation\[2\], MM-estimation\[21\] and Least Trimmed Squares\[12\] which also only require data from a single time point. However, all of these estimation methods do not account for the within-individual correlation of responses.

In this thesis, we propose a new method of estimating the parameters of a random intercept model using only the 2011 Census and IR data. The proposed estimated model can accurately predict the 2016 Census populations of reserves using the 2016 IR data. This parameter estimation method only requires data from a single time point to fit a random intercept model, it is robust to outliers and it takes into account the previous observed response for individuals when predicting.
Chapter 2

Temporal Adjusted Prediction

In this chapter, we propose a new method, temporal adjusted prediction (TAP), for the prediction of responses where data is complete at the first time point, but missing a response at the second time point such as census and Indian Register dataset discussed in the Introduction. The notation used in this thesis is introduced in Section 2.1. Section 2.2 provides details on the motivation behind TAP and explains why linear regression is not appropriate for prediction in datasets where observations on the same set of individuals are taken at different time points. Section 2.3 explains how to predict via linear extrapolation. Section 2.4 covers the details of the temporal adjusted prediction method and introduces a robust variant which can fit the data well even when outliers are present.

2.1 Notation

For this thesis, \( y_{jk}, k \in [1, \cdots, n] \) represents the response at the \( j \)th time point for the \( k \)th individual and \( \hat{y}_{jk} \) is the predicted value of the response at the \( j \)th time point for the \( k \)th individual. In the context of the reserve populations dataset, the term “individual” refers to a specific reserve rather than a person living on a reserve. The value of the \( i \)th covariate/auxiliary variable at the \( j \)th time point for the \( k \)th individual is given by \( x_{ijk}, i \in [1, \cdots, p], k \in [1, \cdots, n] \). For example, \( X_{123} \) are the values of first or primary covariate at the second time point for the third individual in the dataset. In some instances, vector and matrix notation is used as well. The
responses at the \( j \)th time point for \( n \) individuals is denoted \( Y_j = [Y_{j1}, \cdots, Y_{jn}]^T \)
and \( X_j \) is a \( n \times p \) matrix of the covariates/auxiliary variables at the \( j \)th time point for all individuals. The vector notation also applies to the predicted responses \( \hat{Y}_j = [\hat{Y}_{j1}, \cdots, \hat{Y}_{jn}]^T \).

In both the linear regression and random intercept models, \( \beta = [\beta_0, \beta_1, \cdots, \beta_p] \)
is a \( (p + 1) \times 1 \) vector of population parameters that relate \( Y_{jk} \) and \( X_{ijk} \) while \( \hat{\beta} \)
are the estimated values of the parameters \( \beta \). In the random intercept model, \( \beta_0 \) represents the fixed component of the intercept while \( b_{0k} \) represents the random component of the intercept for the \( k \)th individual. In the linear regression model, \( \varepsilon_{jk} \) represents random error at time \( j \) for the \( k \)th individual which we assume to follow a \( \text{Normal}(0, \sigma^2) \) distribution. This differs from the random intercept model where the random error term will be denoted as \( e_{jk} \) which we assume to follow a \( \text{Normal}(0, \phi^2) \) distribution.

### 2.2 Linear Regression

As mentioned in the Introduction, the motivation for temporal adjusted prediction comes from the problem of predicting the populations of Indian reserves in Canada using an auxiliary data source, the Indian Register. The on-reserve population according to the Indian Register is strongly correlated with the reserve population from the census for a given census year. The initial approach was to fit a linear regression model using least squares on past population data gathered from the 2011 IR along with population and geographic data from the 2011 Census, and then use the fitted model to predict the census population of a reserve in 2016 based on data from the IR in 2016.

The linear regression model using data at time point \( j \) is defined as

\[
Y_{jk} = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ijk} + \varepsilon_{jk}, k \in [1, \cdots, n],
\]

where \( Y_{jk} \) is the response, \( X_{ijk} \) are the values of covariates, \( \varepsilon_{jk} \) is the random error for the \( k \)th individual at the \( j \)th time point. The regression coefficients \( \beta_i \) model the effects of each covariate on the response and are assumed to be constant over time. In the context of the Indian reserve populations data, the response variable
\( Y_{jk}, j \in [1, 2, 3], k \in [1, \cdots, n] \) corresponds to the population of the \( k \)th reserve as measured at three time points, the census in 2006, 2011, and 2016. The value of the \( i \)th covariate for the \( k \)th reserve at the \( j \)th time point is given by \( X_{ijk}, j \in [1, 2, 3] \). Covariates include the population of a reserve according to the IR for a given census year and geographical variables such as the province that a reserve is located in.

We fit a linear model using data from the most recent time point with complete data, both response and covariates. We estimate the set of coefficients that results in the minimum sum of squared residuals where the residuals are \( r_j = \hat{Y}_j - Y_j \). The estimated coefficients via least squares estimation are given by \( \hat{\beta}_j = (X_j^T X_j)^{-1} X_j^T Y_j \). In this thesis, unless otherwise specified, the parameters of linear regression models are estimated using least squares.

When predicting with linear regression, we assume that the underlying model between the response and explanatory variables remains constant over time, otherwise it does not make sense to predict using the estimated coefficients \( \hat{\beta}_j \) from one time point on data from another time point. In mathematical terms, we assume that \( \hat{\beta} = \hat{\beta}_1 = \hat{\beta}_2 \) and so on, which is why we drop the subscript \( j \). Predicted values of the response at time point \( j \) are given by \( \hat{Y}_j = X_j^T \hat{\beta} \). Although the linear regression fit is able to capture the linear pattern between the census and IR populations well, the resulting predictions of the reserve populations are poor. Even though the IR on-reserve and census populations are not perfectly correlated, in most cases it is reasonable to expect a proportional increase in the census population if we observe an increase in the IR on-reserve population. In several cases, despite seeing an increase in population on the IR between 2011 and 2016, the model predicted that a reserve would have a lower population in the 2016 Census compared to the 2011 Census.

This is counterintuitive since we expect the growth in a given reserve population to be reflected by both sources. However, it does not make sense to predict a lower population for a reserve, given that we know past census population of the reserve and that we observe a growth in population according to the IR. If we do not have access to the previous census populations for the reserves that we are trying to predict on, then it makes sense to predict using linear regression. This issue is illustrated in the Figure 2.1.
Figure 2.1: Prediction of Selected Observation at Second Time Point via Linear Regression

In Figure 2.1, the diagonal blue line represents the estimated linear regression line based on data simulated at the first time point. Predictions at a given value \( X_{1jk} \) correspond to the height of the blue line on the Y-axis for that \( X_{1jk} \) value. For example, the dashed line on the left represents the \( X_{110} \) value at the first time point for the individual marked by a black triangle and the dotted line to the right represents the \( X_{120} \) value at the second time point for this individual. The difference between the two represents the growth in \( X_1 \) between the second and first time points for this individual. To make a prediction for the point labelled with a black triangle, we use the new value of \( X_{120} \) from the second time point and trace the point that intersects the regression line. The predicted value \( \hat{Y}_{20} \) for this individual is marked by the height of the blue triangle. From this plot, we can see that even though the value of \( X_{1,j0} \) for this individual grows over time and \( X_1 \) and \( Y \) are positively associated across reserves, the predicted value of \( \hat{Y}_{20} \) is less than the original value of \( Y_{10} \). The following sections propose methods to address this issue.
2.3 Linear Extrapolation

A naive but intuitive solution to this problem is to simply extrapolate linearly from data in the first time point to obtain predictions for a second time point by using a primary covariate. In the example of Indian reserve populations in Canada, we can produce a simple prediction of the 2016 Census population by calculating the census to IR reserve population ratio of a reserve based on data from 2011 and then multiplying by the 2016 IR on-reserve population. This process is repeated for each individual reserve. For this section, we refer to the IR on-reserve population as the primary covariate and other explanatory variables as secondary covariates. With linear extrapolation we assume that the ratio of the response $Y_{jk}$ and the primary covariate $X_{1jk}$ is constant over time for each individual. For the $k$th individual, we assume

$$\frac{Y_{1k}}{X_{11k}} = \frac{Y_{2k}}{X_{12k}}.$$ 

Therefore, the predicted value of the response variable for an individual at the second time point using linear extrapolation is given by the equation

$$\hat{Y}_{2k} = \frac{Y_{1k}}{X_{11k}} \times X_{12k}.$$ 

The main benefits of the extrapolation are that it is intuitive and easy to predict with. It uses past data points as a starting point for prediction and extrapolates based on an individual reserve’s census to IR population ratio. A ratio greater than one indicates that there are people living on-reserve that are not listed in the IR. One possible explanation for this is that there is a non-band member living on-reserve and therefore would not be listed in the IR, but would still be counted on the census. Other possible explanations for this discrepancy will be discussed in Chapter 4. This ratio of overall population to registered band members living on-reserve is preserved for any prediction. With linear extrapolation, the prediction line for an individual passes through the origin and the previous observed point $(X_{11k}, Y_{1k})$. Therefore if the relationship between $X_{11k}$ and $Y_{1k}$ is positive, then an increase in $X_{1jk}$ over time results in an increase in the predicted value of $\hat{Y}_{jk}$. 

Figure 2.2: Prediction of Selected Observation at Second Time Point via Extrapolation

In Figure 2.2, the red line represents the predicted value of $Y$ for values of $X_{12k}$ from linear extrapolation for the individual marked by the black triangle. The slope is calculated by taking the ratio of $Y_{10}$ to $X_{110}$ for the selected individual. As with Figure 2.1, the dashed line represents the value of $X_{110}$ and the dotted line to the right represents the value of $X_{120}$ for this individual. The predicted value via linear extrapolation for this individual is given by the height of the red triangle. The height of the blue triangle represents the predictions obtained via linear regression as described in Section 2.2. Contrary to prediction with linear regression, when $X_{1jk}$ increases between time points for an individual, the predicted value for $Y_{jk}$ must increase as well and vice versa.

However, predicting using linear extrapolation has some drawbacks as well. In order to extrapolate and predict accurately, there needs to be a covariate that is highly correlated with the response variable. Although this is true in the context of the Indian reserve populations data, other datasets may not have a predictor that is highly correlated with the response. As the correlation between the predictor and the response variable weakens, prediction accuracy from linear extrapolation
deteriorates. Further, linear extrapolation only allows the use of one covariate even though several may be relevant for the prediction of the response variable. In addition, linear extrapolation cannot make predictions if the observed values of $X_{11k}$ are zero for an individual since it will lead to a division by zero error.

In some sense, linear extrapolation can be thought of as a mixed effects model with a fixed intercept and random slope. Each individual has their own random slope, which is the ratio of $Y_{1k}$ to $X_{11k}$ and a fixed intercept of zero. However, being forced to have a fixed intercept of zero may not be desired in some situations. For the Indian reserves population dataset, this constraint is fine since an IR on-reserve population of zero will also mean a reserve population of zero according to the census for most reserves.

Linear extrapolation does not predict well for individuals that have a large relative increase or decrease in their primary covariate between time points. Although the large increase in the primary covariate is not an issue by itself, if the ratio between the response and covariate in the first time point is not accurate, then the difference between the true response and the prediction is amplified. In the context of our case with the reserve population data, suppose a reserve had a population of 20 people according to the 2011 Census and 5 people according to the 2011 IR. The 15 people included in the census count are non-band members that also live on this reserve. Therefore, this reserve has a census to IR population ratio of 4.

If this reserve were to experience a migration of 95 new registered band members moving onto this reserve between 2011 and 2016, then the predicted number of people living on this reserve according to the census would be $4 \times (95 + 5) = 400$. However, in reality the true population is the 20 people who lived on-reserve in 2011 along with the 95 registered band members who have moved onto the reserve since then. Therefore, the prediction is off by $400 - (20 + 95) = 285$ people because the census to IR reserve population ratio is not constant across time for this reserve.

Cases like this are most common in reserves with small populations. If the IR population is small, then the census to IR population ratio is more likely to be large since the denominator is small. These cases are also more likely to be the ones that see a large relative increase in population between censuses. An increase of 20 people doesn’t affect a reserve with 1000 people on it, but the same increase in
a reserve with 10 people there would be extreme. Violations to the assumption of
a constant census to IR on-reserve population ratio could have negative effects on
prediction results, especially for reserves with a low population according to the
IR. Although linear extrapolation addresses the some of the drawbacks of linear
regression, the combination of issues discussed above suggest that extrapolation
may not yield accurate predictions either.

2.4 Temporal Adjusted Prediction
This thesis proposes a new method for predicting values of a response variable
given auxiliary data measured at the same time point and past data that is used to
construct a model that relates the auxiliary data to the response. Temporal adjusted
prediction combines the strengths of both linear regression and linear extrapolation
while attempting to address the drawbacks of each.

The relationship between the response variable at the first time point $Y_{1k}$ and
its covariates $X_{i1k}$ can be modelled by the regression equation

$$Y_{1k} = \beta_0 + \sum_{i=1}^{p} \beta_i X_{i1k} + \epsilon_{1k},$$

where we assume $\epsilon_{1k} \sim \text{Normal}(0, \sigma^2)$.

Our goal is to predict the values of $Y_{2k}$ given new data $X_{i2k}$. Therefore, we
assume that the relationship between the response and the covariates at the second
time point follows a similar model

$$Y_{2k} = \beta_0 + \sum_{i=1}^{p} \beta_i X_{i2k} + \epsilon_{2k},$$

where we assume $\epsilon_{2k} \sim \text{Normal}(0, \sigma^2)$.

Since we assume that the underlying relationship between the response variable
$Y_{jk}$ and the covariates $X_{ijk}$ do not change over time, i.e. the regression coefficients
$\beta$ are the same in both models, we can combine these equations resulting in

$$Y_{2k} - Y_{1k} = (\beta_0 - \beta_0) + \beta_1 (X_{12k} - X_{11k}) + \beta_1 (X_{22k} - X_{21k}) + \cdots + \beta_p (X_{p2k} - X_{p1k})$$
$$+ (\epsilon_{2k} - \epsilon_{1k}).$$
Therefore, the predicted value for $Y_{2k}$ is given by

$$\hat{Y}_{2k} = Y_{1k} + \sum_{i} \hat{\beta}_i (X_{i2k} - X_{i1k}).$$

(2.1)

The coefficients $\hat{\beta}$ are estimated by least squares on the data from the first time point. If we assume that the relationship between these variables is constant over time then we can apply these estimated coefficients to predict $Y_{2k}$ using $X_{i2k}$. Temporal adjusted prediction is similar to prediction with linear regression in the sense that they share the same set of estimated coefficients. However, the difference between the two is that for linear regression, predictions share a common intercept $\beta_0$ whereas with TAP, the intercept term is random. For TAP, the estimated regression line will interpolate the previous observed response for a given individual.

If complete data is available for more than one time point, then least squares only needs to be fit on the most recent complete data. For example, if we want to predict Indian reserve populations in 2016 and have access to IR on-reserve populations for 2016, 2011, 2006 and census Indian reserve populations for 2011 and 2006, then we only need to estimate the regression coefficients with 2011 data and then use the difference between 2016 and 2011 IR counts along with the 2011 Census count for prediction. Data from 2006 is not used because we have complete data for 2011.

Figure 2.3 shows the prediction mechanisms of all three methods described in this chapter. As with Figure 2.2, the black triangle represents a selected individual that we are interested in making predictions for. The blue and red lines show the estimated linear regression and linear extrapolation lines, respectively. The green line shows the predicted values for this individual using temporal adjusted prediction. Its estimated regression line runs parallel to the linear regression line since both models share the same estimated coefficients, but with an adjustment to the intercept so that it passes through the previous observed point for that individual. The height of the blue and red triangles represent the predictions for this individual at the second time point given by linear regression and extrapolation, respectively. The height of the green triangle represents the prediction given by temporal adjusted prediction.

Temporal adjusted prediction shares many of the same benefits of both methods
and addresses their individual weaknesses. The requirement for the high correlation between the response and primary auxiliary variable is not as important as it is for linear extrapolation since we can include other explanatory variables when fitting the model. These additional covariates may have an effect on the response as well, and these effects can be considered in predictions. Predictions obtained from this method also suffer from less variability compared to those obtained from linear extrapolation since we calculate a common slope for all individuals rather than a slope that only pertains to a single individual. Although this may seem more rigid, fitting a model this way makes more sense since we are able to draw some input from the rest of the dataset rather than just extrapolating based on data from a single individual. In other words, we are assuming that the auxiliary variable(s) have a common effect on the response rather than an individual effect on the response. TAP predictions are more flexible by allowing random intercepts compared to linear regression which assumes a common intercept $\beta_0$. In situations where a large proportion of the error is inherent to specific individuals, TAP is able to account for this and gives better predictions than linear regression which assumes that all
of the error is random. This is further discussed in Subsection 2.4.1.

### 2.4.1 Random Intercept

Data at time point $j$ for $k$ individuals can be modelled by the formula

$$Y_{jk} = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ijk} + \epsilon_{jk}. \quad (2.2)$$

In the linear regression framework, the error term $\epsilon_{jk}$ is assumed to follow a $\text{Normal}(0, \sigma^2)$ distribution. The error terms are also assumed to be independent and identically distributed.

However, when we look at data over more than one time point for the same set of individuals, some of the error terms can no longer be assumed independent. More specifically, the error term $\epsilon_{jk}, k \in [1, \cdots, n]$ can be decomposed into two components, $b_{0k}, k \in [1, \cdots, n]$, a random intercept to model the within individual variation, and $e_{jk}, k \in [1, \cdots, n]$, the remaining random error in the model. The random intercept $b_{0k}$ is combined with the fixed component of the intercept $\beta_0$ to get the intercept term for the random intercept model. For example, in the case of prediction of populations of Indian reserves, suppose a group of individuals living on a particular reserve are not found in the IR but are included in the census count. Although this difference may be considered as error when we fit a linear regression model on the data at the first time point, it can still be considered to predict the population of the reserve at the second time point since it is reasonable to expect that this group will still be living on a particular reserve during the next census. To model the random effects of specific reserves, a random intercept term can be included in the regression model. This random intercept, $b_{0k}, k \in [1, \cdots, n]$, is assumed to follow a $\text{Normal}(0, \tau^2)$ distribution. The remaining portion of the random error that is not explained by the random intercept term is denoted as $e_{jk}$ which is assumed to follow a $\text{Normal}(0, \phi^2)$ distribution. We assume the random intercept term $b_{0k}$ and the random error term $e_{jk}$ are independent of each other.

The typical random intercept model with the two sources of variation at time point $j$ is given by

$$Y_{jk} = (\beta_0 + b_{0k}) + \beta_1 X_{1jk} + \beta_2 X_{2jk} + \cdots + \beta_p X_{pjk} + e_{jk}. $$
The overall variability in the data is a combination of the random intercept and the random error term. Therefore, if $\text{Var}(b_{0k}) = \tau^2$ and $\text{Var}(e_{jk}) = \phi^2$, then the total variance in the random effects model is $\tau^2 + \phi^2$ since we assume that the two sources of variation are independent of each other. The variance of the error term in the linear regression model is equal to the combined variance of the error terms from the random intercept model

$$\sigma^2 = \tau^2 + \phi^2.$$ 

However, the variance of the random intercept cannot be directly estimated if we lack data from at least two time points. TAP addresses this limitation using the response at the first time point instead of the random intercept to predict the response at the second time point.

### 2.4.2 Comparing the MSPE of Linear Regression and Temporal Adjusted Prediction

Since we are interested in comparing the prediction performance of each method, we compare the mean squared prediction error (MSPE) of both linear regression and TAP. The MSPE is given by $\text{MSPE} = \frac{1}{n} \sum(Y_{2k} - \hat{Y}_{2k})^2$ where $Y_{2k}$ and $\hat{Y}_{2k}$ are the observed and predicted response for the $k$th individual at the second time point. In this subsection, we examine the univariate case, although the results can be generalized to include additional covariates.

Since linear regression does not include a random intercept term, the responses $Y_{1k}$ and $Y_{2k}$ can be modelled by

$$Y_{jk} = \beta_0 + \beta_1 X_{1jk} + \epsilon_{jk}, \quad \text{(2.3)}$$

and used to predict the response at the second time point

$$\hat{Y}_{2k} = \hat{\beta}_0 + \hat{\beta}_1 X_{12k},$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated with least squares on data from the first time point. The prediction error is obtained by taking the difference between the predicted response and the actual response.
\[ PE_{LR,k} = \hat{Y}_{2k} - Y_{2k} = (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)X_{12k} - \varepsilon_{2k}. \]

The expected MSPE is given by the variance of the prediction error of linear regression which is

\[
E[MSPE_{LR,k}] = Var(PE_{LR,k}) \\
= Var(\hat{Y}_{2k} - Y_{2k}) \\
= Var((\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)X_{11k} - \varepsilon_{2k}) \\
= Var(\hat{Y}_{2k}) + Var(\varepsilon_{2k}) \\
= \sigma^2 \left\{ \frac{1}{n} + \frac{X_{12k} - \bar{X}_{11k}}{(n-1)s_{X_{11k}}^2} \right\} + \sigma^2.
\]

As \( n \to \infty \), the first term in the MSPE goes to zero and the expected MSPE converges to \( \sigma^2 \).

For TAP, the predicted response at the second time point, \( \hat{Y}_{2k} \) is

\[ \hat{Y}_{2k} = Y_{1k} + \hat{\beta}_1 (X_{12k} - X_{11k}), \]

where \( \hat{\beta}_1 \) is estimated via least squares on data from the first time point. The prediction error of TAP is obtained by

\[ PE_{TAP,k} = \hat{Y}_{2k} - Y_{2k} = Y_{1k} - Y_{2k} + \hat{\beta}_1 (X_{12k} - X_{11k}). \]

The expected MSPE is given by the variance of the prediction error of TAP which is
\[ E[\text{MSPE}_{\text{TAP},k}] = \text{Var}(PE_{\text{TAP},k}) \]
\[ = \text{Var}(\hat{Y}_{2k} - Y_{2k}) \]
\[ = \text{Var}(Y_{1k} - Y_{2k} + \hat{\beta}_1 (X_{12k} - X_{11k})) \]
\[ = \text{Var}(Y_{1k} - Y_{2k}) + \text{Var}(\hat{\beta}_1) \cdot (X_{12k} - X_{11k})^2 \]
\[ = 2\phi^2 + \frac{(\phi^2)}{(n-1)s_{122}^2} \cdot (X_{12k} - X_{11k})^2. \]

As \( n \to \infty \), the second term in the expected MSPE goes to zero and the expected MSPE converges to \( 2\phi^2 \).

Since the expected MSPE of TAP converges to \( \phi^2 \) and the expected MSPE of linear regression converges to \( \sigma^2 = \phi^2 + \tau^2 \), the method that will predict more accurately depends on the underlying parameters \( \phi \) and \( \tau \) as \( n \to \infty \). In cases where all of the variation is random error, \( \tau^2 = 0 \), then linear regression should have an expected MSPE that is half of the expected MSPE of TAP when \( n \) is large since the expected \( \text{MSPE}_{\text{TAP},k} \to 2\phi^2 \) while \( \text{MSPE}_{\text{LR},k} \to \sigma^2 = \phi^2 + \tau^2 = \phi^2 \). If the variance of the random error term is equal to the variance of the random intercept, \( \phi^2 = \tau^2 \), then the expected MSPE of each method converges to the same value since \( \text{MSPE}_{\text{TAP},k} \to 2\phi^2 \) while \( \text{MSPE}_{\text{LR},k} \to \sigma^2 = \phi^2 + \tau^2 = 2\phi^2 \).

From this, we can conclude that if \( \phi < \tau \), then \( E[\text{MSPE}_{\text{TAP},k}] < E[\text{MSPE}_{\text{LR},k}] \) as \( n \to \infty \). In other words, when the variance of the random intercept is larger than the variance from the random error term, we expect TAP to predict better than linear regression.

Figure 2.4 and Table 2.1 compare the MSPE given by linear regression and temporal adjusted prediction for different values of \( \tau \) and fixed \( \phi = 5 \). For this simulation, datasets of \( n=500 \) observations are generated with a single covariate with values of responses and covariates from two time points. The random intercept \( b_{0k} \) is generated for each trial from a Normal\((0, \tau^2)\) distribution where \( \tau \in [0, 10, 20, 30, 40] \). The results of 1000 independent trials are shown. More specific details about the simulation settings are described in Section 3.1.

Figure 2.4 and Table 2.1 show that linear regression performs better than temporal adjusted prediction when \( \tau = 0 \), but as \( \tau \) increases, then temporal adjusted
Figure 2.4: Mean Square Prediction Error of Temporal Adjusted Data Using Temporal Adjusted Prediction and Linear Regression

<table>
<thead>
<tr>
<th>Method</th>
<th>τ = 0</th>
<th>τ = 10</th>
<th>τ = 20</th>
<th>τ = 30</th>
<th>τ = 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal Adjusted Prediction</td>
<td>50.04</td>
<td>50.06</td>
<td>50.06</td>
<td>50.11</td>
<td>50.41</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>25.15</td>
<td>124.43</td>
<td>422.88</td>
<td>920.93</td>
<td>1623.87</td>
</tr>
</tbody>
</table>

Table 2.1: Mean Squared Prediction Error of Temporal Adjusted Prediction and Linear Regression when varying Standard Deviation of Intercept

prediction predicts $Y_{2k}$ more accurately than linear regression.

These results are not unexpected since when $\tau = 0$, there is no random intercept and this means that almost all of the data in both time points line up well with the estimated regression line. The only variability remaining in the model is produced by $e_{1k}$ and $e_{2k}$. If the linear regression model fits well, then the prediction error is due to the inability of the fitted linear regression to predict $e_{2k}$. However, since the temporal adjusted prediction includes $Y_{1k}$ in its predictions, it includes $e_{1k}$ into its predictions as well. Combined with the error from $e_{2k}$, this explains why the MSPE for temporal adjusted prediction is roughly double that of linear regression when there is no variation in the random intercept.

The MSPEs of TAP and linear regression are in line with the theoretical expected MSPEs calculated above. In these simulations, $\phi^2 = 25$ and therefore the
MSPE of TAP for each value of $\tau$ is approximately $2\phi^2 = 50$. The observed MSPE of linear regression is approximately $\phi^2 + \tau^2 = 25 + [0, 10^2, 20^2, 30^2, 40^2]$ when $\tau = [0, 10, 20, 30, 40]$.

### 2.4.3 Robust Temporal Adjusted Prediction

It is well documented that least squares estimates are highly sensitive to outliers [21]. Least squares estimates are obtained by minimizing the total sum of squared residuals. Even one extreme point in the dataset can result in a large residual and an even larger squared residual. In order to minimize the sum of squared residuals, the least squares regression line will be skewed towards this outlier. This effectively causes the outlier to have more influence on the estimated regression coefficients than non-outlying points. In order to provide additional robustness to outliers, estimated regression coefficients used in the temporal adjusted prediction model are estimated using a robust procedure rather than being estimated by least squares.

In our motivating example, when we plot the relationship between the Indian reserve populations according to the 2011 Census and the Indian Register in 2011, we can see that the relationship between these two variables is positive and strongly linear. However, in Figure 2.5, we can see that some potential outliers exist in the dataset. These bands have low reserve populations according to the IR but high reserve populations according to the census. Possible causes of these outliers are discussed in Chapter 4, but they present a problem when it comes to estimating regression coefficients using least squares.

Although it is possible to identify and remove outliers from the dataset manually in simple cases where outliers are obvious, classifying an observation as an outlier is not always trivial. A point that may look extreme can possibly be explained by another variable and may not be an outlier once additional factors are considered. Some points may not be obvious outliers but can still be influential. In Figure 2.5, some potential outliers are highlighted in blue. Some of these observations are obvious outliers since they are located far away from the bulk of the data. However, for others it is not clear whether these observations are true outliers or if they occur due to natural variability.

There are many robust regression methods in the current literature but for tem-
poral adjusted prediction, we estimate the regression parameters $\beta$ using MM-Estimation introduced by Yohai (1987) [21]. In robust statistics, a common measure of an estimators’ resistance to outliers or robustness is the breakdown point introduced by Hampel (1971) [1]. The breakdown point of an estimator is the largest proportion of outliers a sample can contain before the estimated parameters are broken [21]. The main benefits of using MM-estimation are that it has a high breakdown point and is highly efficient when the errors are normally distributed [21].

In order to define MM-estimator, we must first define scale M-estimators [2][21]. As stated by Yohai (1987) [21], let $\rho$ be a real function that satisfies the following assumptions:

1. $\rho(0) = 0$
2. $\rho(-u) = \rho(u)$
3. $0 \leq u \leq v \implies \rho(u) \leq \rho(v)$
4. $\rho$ is continuous
5. Let $a = \sup \rho(u)$, then $0 < a < \infty$

6. If $\rho(u) < a$ and $0 \leq u \leq v$, then $\rho(u) < \rho(v)$

The M-scale estimator $s(u)$ of a sample $u = (u_1, u_2, \ldots, u_n)$ is the value $s$ that satisfies

$$\frac{1}{n} \sum_{i=1}^{n} \rho(u_i/s) = b,$$

(2.4)

where $b = E_{\phi}(\rho(u))$ and $\phi$ is the standard normal distribution [21].

The MM-estimator is calculated using the following three-step procedure. First, take an initial estimate $T_{0,n}$ of the regression coefficients $\beta$ with a high breakdown point. Next, compute the residuals $r(T_{0,n})$ and compute the scale M-estimate of the residuals $s_n = s(r(T_{0,n}))$ as in equation 2.4 such that $a = \max \rho_0(u)$ and $b/a = 0.5$. Lastly, let $\rho_1$ be function where $\rho_1(u) \leq \rho_0(u)$, $\sup \rho_1(u) = \sup \rho_0(u) = a$ and $\psi_1 = \rho_1'$. The MM-estimate $T_{1,n}$ is the solution of

$$\sum_{i=1}^{n} \psi_1(r_i(\beta)/s_n)x_i = 0,$$

(2.5)

which verifies $S(T_{1,n}) \leq S(T_{0,n})$ where

$$S(\beta) = \sum_{i=1}^{n} \rho_1(r_i(\beta)/s_n) = 0,$$

(2.6)

and $\rho_1(0/0)$ is defined as 0 [21].

In order to apply MM-estimates to temporal adjusted prediction, we replace the estimated coefficients via least squares $\hat{\beta}$ in equation 2.1 with those estimated via MM-estimation. This is referred to as robust temporal adjusted prediction, or robust TAP. For this thesis, MM-estimates are computed in R using the package robustbase [11].
Chapter 3

Simulation Study

In this chapter, we compare the prediction performance of temporal adjusted prediction and robust temporal adjusted prediction against those of other selected models including linear regression, robust regression and extrapolation under a variety of simulation settings. Section 3.1 describes the method of data generation for the simulations in this chapter which is based on the Indian reserve populations dataset mentioned throughout this thesis. Features present in that dataset (including outliers, additional covariates and data from an additional time point) are individually added to the simulations so that we can observe their effects on the prediction accuracy for each of the models listed above.

Section 3.2 shows the prediction performance, measured by the MSPE, of the selected models on our simulated data. Subsection 3.2.1 shows the results of simulations carried out under our default settings without the inclusion of any additional features. In Subsection 3.2.2, we add outliers into the dataset and then compare the prediction performance of each model. Subsection 3.2.3 discusses the effects of adding additional covariates on the prediction accuracy of each model. Lastly, in Subsection 3.2.4, we add data from a third time point into the dataset to accommodate the inclusion of a linear mixed effects model with a random intercept and fixed slopes, and then discuss the prediction performance of each model. The parameters of this linear mixed effects model are estimated using maximum likelihood methods.
3.1 Data Generation

This section explains how data is generated for the simulations performed in this chapter. The data is generated according to the random intercept model described in Section 2.4.1. Specific values of parameters and distributions of independent covariates are meant to model the Indian reserve populations dataset described in Chapter 4. For example, in the Indian reserve populations dataset, the IR on-reserve population roughly follows an exponential distribution, so the distribution of the covariate \( X_{1jk} \) in the data generation model also follows an exponential distribution. This section only describes the default settings of data generation that are common across all simulations. Differences in data generation methods that are specific to a certain setting are described in the corresponding subsection.

For simulations in Subsection 3.2.1, we generate a dataset of \( n = 500 \) observations for two time points according to a random intercept model with a single covariate and a random error term. The data for the first time point is generated according to the formula

\[
Y_{1k} = (\beta_0 + b_{0k}) + \beta_1 X_{11k} + e_{1k},
\]

where \( X_{11k} \sim \text{Exp}(0.001) \), \( \beta_0 = 0 \), \( \beta_1 = 0.82 \) and \( e_{1k} \sim \text{Normal}(0, \phi^2) \) with \( \phi = 10 \). The random intercept term \( b_{0k} \) follows a \( \text{Normal}(0, \tau^2) \) distribution with \( \tau = 20 \). It is important to note that the values of \( b_{0k} \) are exactly the same for both \( Y_{1k} \) and \( Y_{2k} \) since this random term models the variability that is inherent to a specific individual across both time points.

Since data for the second time point must be correlated with the first time point, we multiply the independent \( X_{11k} \) variable by a growth factor \( \alpha_1 = 1.2 \) and then add some random noise \( \eta_{1k} \sim \text{Normal}(0, 10^2) \) to get \( X_{12k} \). This growth factor is meant to emulate the growth in population that reserves will experience between censuses. The value of \( X_{12k} \) is given by \( X_{12k} = \alpha_1 X_{11k} + \eta_{1k} \). Although this growth factor is only applied to \( X_{11k} \), growth in the response variable \( Y_{2k} \) will be reflected when \( Y_{2k} \) is calculated according to the formula

\[
Y_{2k} = (\beta_0 + b_{0k}) + \beta_1 X_{12k} + e_{2k}.
\]
The response at the second time point $Y_{2k}$ is generated using the same values of $\beta_0, b_{0k}, \beta_1$ values used to generate $Y_{1k}$. The random error term $e_{2k} \sim \text{Normal}(0, \phi^2)$ with $\phi = 10$. The random error terms at the first and second time points, $e_{1k}$ and $e_{2k}$, are identically and independently distributed.

All results shown in this chapter are based on 1000 independent trials in each setting.

### 3.2 Simulation Results

#### 3.2.1 Default Settings

The MSPE of 1000 independent trials under the default simulation settings are shown in the boxplots in Figure 3.1 with the median MSPEs reported in Table 3.1. A few extreme values of MSPE for extrapolation are not shown for some plots because they would distort the Y-axis of the plot if included. However, they are still factored into the median MSPEs shown in the tables. The number of omitted values for each plot is mentioned in the corresponding subsection. Some of the distributions of MSPE are right skewed so the median is a better summary statistic of different methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extrapolation</td>
<td>105.43</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>421.61</td>
</tr>
<tr>
<td>Robust Regression</td>
<td>421.59</td>
</tr>
<tr>
<td>Robust Temporal Adjusted Prediction</td>
<td>49.95</td>
</tr>
<tr>
<td>Temporal Adjusted Prediction</td>
<td>49.93</td>
</tr>
</tbody>
</table>

**Table 3.1:** Comparison of Median MSPE under Default Settings

Table 3.1 shows that TAP and robust TAP achieve the lowest median MSPEs followed by extrapolation and then by linear regression and robust regression. This result is expected since it was already shown that TAP predicts better than linear regression when the variance of the random intercept term is large relative to the variance of the random error term. Extrapolation usually achieves a MSPE somewhere between linear regression and TAP.
It should be noted that 92 MSPE points for the extrapolation method are not displayed in the Figure 3.1 since they exceed a maximum bound of 800 and including all values would distort the Y-axis in the plot. This is likely due to the fact that points with a very low $X_{11k}$ value can result in a very high $\frac{Y}{X_{11k}}$ ratio. If $X_{11k}$ is very small for an individual and it experiences even a moderate increase, then the prediction for that individual will be extremely poor. Even a single bad prediction can badly affect the MSPE which occurs in a number of trials in this simulation.

### 3.2.2 Outliers

In this section, we show the effects of adding outliers to the generated dataset on prediction performance. Some outliers are found in the Indian reserve populations dataset and the outliers generated in this section are meant to mimic their location.
Figure 3.2: Comparison of MSPE under Default Settings (Zoomed In)

and effects on the fitted models. In the first simulation, we randomly replace 5% of the data with points that are generated following

\[ X_{11k} \sim Exp(0.01), \]
\[ Y_{1k} = (\beta_0 + b_{0k}) + \beta_1 X_{11k} + e_{1k}. \]

where \( \beta_0 = 0, \beta_1 = 0.82 \) and \( e_{1k} \sim Normal(0, \phi^2) \) with \( \phi = 10 \). Note that the \( X_{11k} \) values for the outliers are generated in a similar fashion to those from non-outliers with the exception that the rate parameter of the exponential distribution changes from 0.001 to 0.01. This adjustment is made to reflect the fact that most outliers located in the reserve populations dataset have lower values IR on-reserve populations, but higher census populations for a given year. To match the
higher values of the response for the outliers in this simulation, we generate the random intercept term from $b_{0k}$ from a non-standardized Student’s t-distribution $t_{d.f. = 1}(\mu = 1000, \sigma = 10)$. Since the generated dataset focuses on the same set of individuals over time, the outliers are also present in $Y_{2k}$ since $Y_{1k}$ and $Y_{2k}$ are both generated using the same random intercept, $b_{0k}$.

The MSPEs of 1000 trials are shown in the boxplots in Figure 3.3. The corresponding median MSPEs are displayed in Table 3.2.

![Figure 3.3: Comparison of Mean Squared Prediction Error with 5% Outliers](image)

When the data contains outliers, the prediction performance of TAP and robust TAP is approximately the same as when it contained no outliers while all other methods experience large increases in their MSPE. Once again, this behaviour is expected since, a robust estimation method should still be able to accurately estimate the true regression parameters when there is a small amount of outlier
**Figure 3.4:** Comparison of Mean Squared Prediction Error with 5% Outliers (Zoomed In)

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extrapolation</td>
<td>32491.85</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>46411.11</td>
</tr>
<tr>
<td>Robust Regression</td>
<td>50618.20</td>
</tr>
<tr>
<td>Robust Temporal Adjusted Prediction</td>
<td>50.05</td>
</tr>
<tr>
<td>Temporal Adjusted Prediction</td>
<td>195.46</td>
</tr>
</tbody>
</table>

**Table 3.2:** Comparison of Median MSPE with 5% Outliers

contamination, while the least squares estimates may be more severely affected. It may seem counter-intuitive that robust regression predicts worse than linear regression when outliers are included but this behaviour is expected. Robust regression
methods are able to estimate the underlying distribution of the non-contaminated points while placing less weight on outliers whereas linear regression using least squares weighs the outliers more heavily than the non-contaminated data in this simulation. When it comes to prediction, robust regression can predict the non-contaminated data well but poorly predicts the outliers since it is not trying to fit those points. Linear regression does a mediocre job at fitting the outliers and the non-contaminated points and does a mediocre job predicting both as well. As a result, robust regression does a better job predicting the non-contaminated points, but linear regression does better on predicting the overall dataset. Other robust measures of prediction performance such as median absolute prediction error (MAPE) are a suitable measure of performance prediction when outliers are present. This performance measure will be used in Chapter 4 along with MSPE.

The vertical displacement of the outliers are included in the random intercept term \( b_{0k} \) and therefore do not affect the underlying relationship between \( X_{1jk} \) and \( Y_{1k} \), although they may affect how well each method is able to estimate \( \beta_1 \). This is why TAP is still able to predict relatively well when outliers are present. A point that is in an extreme location in the first time point will also be in an extreme location in the second time point. If there is little growth between the first and second time points in those outlying points in \( X_{1jk} \) and \( Y_{jk} \), then the value of the first time point \( Y_{1k} \) should approximate \( Y_{2k} \) well even if the linear regression fit is not able to accurately estimate \( \beta_1 \). Although TAP is able to predict relatively well when outliers are present, there are gains to be made by using its robust alternative instead since robust estimation methods are able to estimate \( \beta_1 \) better and therefore predict more accurately. By starting the predictions at an already outlying point \( Y_{1k} \), robust TAP overcomes the drawback of not being able to predict the outliers that robust regression suffers from. Also, note that 271 extreme values of MSPE are omitted from Figure 3.3 since their inclusion would distort the axes in Figure 3.3.

In Figure 3.5 we show the prediction results when the proportion of outliers in the simulated data is increased to 10%. As more outliers are introduced into the data, the results discussed above become even more defined. Robust TAP has a similar MSPE compared to when the data contains only 5% outliers instead of 10%. Other models show substantial increases in MSPE when the proportion of outliers is increased. TAP is the next best performing model in terms of median.
Figure 3.5: Comparison of Mean Squared Prediction Error with 10% Outliers

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extrapolation</td>
<td>131135.92</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>84931.00</td>
</tr>
<tr>
<td>Robust Regression</td>
<td>101153.51</td>
</tr>
<tr>
<td>Robust Temporal Adjusted Prediction</td>
<td>49.90</td>
</tr>
<tr>
<td>Temporal Adjusted Prediction</td>
<td>557.25</td>
</tr>
</tbody>
</table>

Table 3.3: Comparison of Median MSPE with 10% Outliers

MSPE followed by linear and robust regression and then extrapolation.
3.2.3 Additional Covariates

In this simulation, we examine the effects of adding two additional covariates into the model, one quantitative and one binary. In addition to making highly variable predictions when the values of $X_{11k}$ are low, another major drawback of linear extrapolation is that it cannot easily factor in the effects of additional covariates to make predictions and therefore it will not be featured in this section. The response variable at the first time point, $Y_{1k}$, is generated according to

$$Y_{1k} = (\beta_0 + b_{0k}) + \beta_1 X_{11k} + \beta_2 X_{21k} + \beta_3 X_{31k} + e_{1k},$$

where $X_{21k} \sim \text{Normal}(0, 20^2)$ and $X_{31k} \sim \text{Bernoulli}(0.5)$ with the corresponding
coefficients $\beta_2 = -20$ and $\beta_3 = 100$. Other parameters and variables, $\beta_1, X_{11k}, e_{1k}$ and $b_{0k}$, are still created according to the default settings described in Section 3.2.1. The second covariate at the second time point $X_{22k}$ is generated in a similar manner as $X_{12k}$. A growth factor $\alpha_2 = 1.3$ is applied to $X_{21k}$ and then some random error $\eta_{2k} \sim \text{Normal}(0, 10^2)$ is added

$$X_{22k} = \alpha_2 \ast X_{21k} + \eta_{2k}.$$ 

Lastly, $X_{31k}$ does not change over time, hence $X_{31k} = X_{32k}$.

**Figure 3.7:** Comparison of Mean Squared Prediction Error with an Additional Variable

The results shown in Figure 3.7 and Table 3.4 are comparable to those in Subsection 3.2.1. In fact, the median MSPEs shown in Table 3.4 are almost identical to
Table 3.4: Comparison of Median MSPE with Additional Covariates

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>422.71</td>
</tr>
<tr>
<td>Robust Regression</td>
<td>423.61</td>
</tr>
<tr>
<td>Robust Temporal Adjusted Prediction</td>
<td>50.134</td>
</tr>
<tr>
<td>Temporal Adjusted Prediction</td>
<td>50.10</td>
</tr>
</tbody>
</table>

those shown in Table 3.1 which reports the median MSPE under the default simulation settings. All models are able to estimate the additional regression coefficients $\beta_2$ and $\beta_3$ well and therefore accurately predict the effects that $X_{22}$ and $X_{32}$ have on the response $Y_2$. The increase in prediction error comes from the additional vari-
ability introduced by $\eta_{2k}$. From this simulation, we can conclude that none of the included estimation procedures are heavily impacted by the inclusion of additional covariates.

### 3.2.4 Additional Time Point

In this subsection, we add an additional time point to the dataset to accommodate the inclusion of a linear mixed effects model which is more suitable for fitting longitudinal data compared to linear regression. Although temporal adjusted prediction can be considered a special case of linear mixed effects models, when we mention LME models in this thesis, we are referring to LME models where the parameters are estimated using maximum likelihood methods. The parameters of TAP are estimated as described in Section 2.4. Since we are introducing an additional time point, we need to define notation that is specific to this subsection. Most parameters and variables take on the same meaning as described in Section 3.1 with the exception of a new term $\beta_{1j}$. The parameter $\beta_{1j}$ represents the true value of the regression coefficient $\beta_1$ at time point $j$. The data generation model for the response variable $Y_{jk}$ at time point $j \in [1,2,3]$ is given by

$$Y_{jk} = (\beta_0 + b_{0k}) + \beta_{1j}X_{1jk} + e_{jk},$$

where $X_{11k}$ and $X_{12k}$ are generated the same way as described in the default settings. $X_{13k}$ is generated by multiplying $X_{12k}$ by a growth factor $\alpha_1 = 1.2$ and then adding random error $\eta_{2k} \sim \text{Normal}(0,10^2)$. The random error term $e_{jk} \sim \text{Normal}(0,5^2)$ at each time point $j$.

In this subsection, we also examine the effects of violations to our constant slope assumption. In this section, we denote $\beta_{1j}$ as the true regression coefficient that represents the relationship between covariate $X_{1jk}$ and response $Y_{jk}$ at time point $j \in [1,2,3]$. In previous subsections, we assumed that $\beta_{1j}$ remains constant over time; however in practice this may not always be the case. The constant relationship over time is one of the key assumptions for predicting with TAP and LME models. We investigate the effects of slight deviations of this assumption on prediction accuracy.

In the case of the reserve populations dataset, fitting a linear regression model
at each of the available time points shows that there are small differences in the estimated coefficient $\beta_{1j}$ at each time point and therefore the assumption of a constant relationship between $X_{1jk}$ and $Y_{jk}$ may be not be true. We test our selected models under three different scenarios to see the effects of violations to the assumption that $\beta_1$ is constant. Model parameters are estimated with complete data from the first and second time points available, and used to predict $\hat{Y}_{3k}$ given $X_{13k}$.

**Scenario (1):** $\beta_1$ is constant over time, $\beta_{11} = \beta_{12} = \beta_{13} = 0.82$

This is the default setting where the relationship between X and Y is constant over time.

**Scenario (2):** $\beta_1$ is not constant over time, $\beta_{11} \neq \beta_{12} = \beta_{13}$

In this scenario, we test to see the effects of changing $\beta_{11}$ to 0.84 while the other two coefficients $\beta_{12}$ and $\beta_{13}$ remain at 0.82. This modification should only affect the linear mixed effects model since it relies on data from first and second time points to fit the model, whereas the other models only utilize data from the second time point. We test this to see how a small change in $\beta_{11}$ can affect the predictions produced by the linear mixed effects model.

**Scenario (3):** $\beta_1$ is not constant over time, $\beta_{11} = \beta_{13} \neq \beta_{12}$

This is similar to the second scenario except here we change $\beta_{12}$ to 0.84 while the other two coefficients $\beta_{11}$ and $\beta_{13}$ remain at the default value of 0.82. Since all model parameters are estimated using data at the second time point, a change to the coefficient $\beta_{12}$ should negatively impact the prediction accuracy of all three models. The fitted models are trying to predict data points that are generated by slightly different parameter values.

In this subsection, we fit a linear mixed effects model with a random intercept, temporal adjusted prediction, robust temporal adjusted prediction, linear regression, robust regression and extrapolation models.

The results in Figure 3.9 and Table 3.5 show that in **Scenario (1)** when the slope $\beta_1$ is constant over time, the LME model gives the best predictions followed closely by robust TAP, TAP and then linear extrapolation. Linear regression and robust regression are not shown in Figure 3.9 so that their inclusion does not distort the axes.

In **Scenario (2)** we see that even a small change in one of the coefficients has a noticeable effect on the prediction ability of the LME model. The LME model
Figure 3.9: Comparison of Mean Squared Prediction Error with Data with an Additional Time Point: Scenario (1)

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPE (1)</th>
<th>MSPE (2)</th>
<th>MSPE (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extrapolation</td>
<td>99.23</td>
<td>99.23</td>
<td>1796.79</td>
</tr>
<tr>
<td>Linear Mixed Effects Model</td>
<td>37.70</td>
<td>990.14</td>
<td>5152.61</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>424.38</td>
<td>424.38</td>
<td>2076.60</td>
</tr>
<tr>
<td>Robust Regression</td>
<td>424.30</td>
<td>424.30</td>
<td>2078.29</td>
</tr>
<tr>
<td>Robust Temporal Adjusted Prediction</td>
<td>49.93</td>
<td>49.93</td>
<td>1702.12</td>
</tr>
<tr>
<td>Temporal Adjusted Prediction</td>
<td>49.94</td>
<td>49.94</td>
<td>1702.06</td>
</tr>
</tbody>
</table>

Table 3.5: Comparison of Median MSPE with Data with an Additional Time Point under three Scenarios

goes from having the lowest median MSPE in the previous scenario to having the
highest. The other models have exactly the same median MSPE because their parameters are estimated using only data from the second time point so a change in $\beta_{11}$ does not affect them.

When $\beta_{12}$ is modified as in Scenario (3), the prediction accuracy of all models decreases. This is expected since the parameters of all models except for the LME model are estimated using data exclusively from the second time point. If this data is generated with a different slope $\beta_{12}$ at 0.84, then the fitted models are not able to predict the points from the third time point well since the model fitted on the data at the second time point is not the same as the underlying model that generated the data at the third time point. However, in this scenario the LME model still predicts worse than the other methods.
Figure 3.11: Comparison of Mean Squared Prediction Error with Data with an Additional Time Point: Scenario (3)

From these results, we conclude that the LME model performs the best when the coefficient $\beta_{1j}$ is constant over time. However, even a small change in the coefficient between time points can have damaging effects on prediction accuracy for the LME model. Although the prediction accuracy of other models is also affected when $\beta_{1j}$ changes between time points, they do not suffer to the same extent as the LME model. From these results, we can see that the LME model is more sensitive to violations of the assumption of a constant relationship between the covariates and the response variable, compared to TAP. In the Indian reserve populations dataset, we expect there to be small changes in the regression parameters between censuses. In the exploratory analysis, when robust regression models are fitted using data from 2006 and 2011, we see that there are small differences in the
regression parameters between the two time points. This will be further discussed in Section 4.3.1.
Chapter 4

Case Study: Prediction of Populations of Indians Reserves in Canada

As discussed during the Introduction, the main motivation behind the development of temporal adjusted prediction comes from the goal of predicting the population of Indian reserves in Canada for the 2016 Census using the Indian Register (IR) as an auxiliary data source. These predictions are used to confront the actual counts of the 2016 Census. If there are large discrepancies between the actual counts and model expectations, then it could warrant further investigation to determine the cause of the difference. Most reserve inhabitants are band members that are registered on the IR, [20] [16] so the IR should provide up to date information on the population of each reserve [6]. If there are changes to the population of a reserve according to the IR, then these changes are expected to be reflected by the census as well. In Canada, the census is conducted every five years, on years ending with a 1 or 6, since 1871. However, data from the IR on band populations is only available starting from 2006 so we can only use data from the 2006, 2011 and 2016 Censuses and IR for this case study. Using the IR and census data from 2006 and 2011, we construct a model that can predict the populations of reserves for 2016 based on population counts from the 2016 IR. In this chapter, we take a deeper look at the data from the 2006, 2011 and 2016 Censuses and the Indian Register.
and compare the prediction accuracy of temporal adjusted prediction against the prediction accuracy of the other methods discussed in this thesis. In Section 4.1, we introduce some background information regarding the census and IR necessary for the understanding of this case study. Section 4.2 provides a more in-depth explanation of the data used in this study and some of the steps required to create a dataset ready for analysis. Lastly, Section 4.3 describes the analysis of the Indian Register and census data. This section also compares the prediction performances results of temporal adjusted prediction and other models when used to predict the population of Indian reserves in Canada for the 2016 Census.

4.1 Background

4.1.1 The Indian Register

The Indian Register is the official record of all Registered Indians in Canada. The IR was created in 1951 by the government in order to identify individual First Nations people and the bands to which they belonged [5]. This information is used to determine to which benefits people are entitled through treaty agreements with bands. In order for a person to become a Registered Indian, they have to prove their descent from a person who is also a Registered Indian. Additions and deletions to the IR are made by the Indian Registrar, an employee of Indigenous and Northern Affairs Canada (INAC). The Indian Registrar is responsible for determining whether or not a person is eligible for registration on the IR [5].

Although the IR is a relational database that was originally created for administrative purposes, the information that it contains can still be useful for statistical analysis and modelling. Based on the information available from the IR, we can obtain the number of registered band members living on a band’s own reserves. The variables that are necessary for calculation of own-band on-reserve population counts include the band membership, the location of residence (whether an individual lives on their own bands reserve or not), and their activity status. Most individuals on the IR are active; individuals who are not active include those who are presumed or confirmed as deceased, and those who have given up their Registered Indian status. The IR only states whether an individual lives on a reserve
that belongs to their own band or not and does not differentiate between specific reserves if a band is associated with multiple reserves. The IR also indicates if an individual lives on a reserve of another band, but since it does not list which other specific band or reserve, this information is not used. The in-scope population from the Indian Register are individuals who are currently indicated as living on the reserve of their own band and are also currently active. The data used in this thesis captures the state of the Indian Register on May 31st of each census year and is tabulated to produce counts of registered band members on each band’s own reserves. The IR on-reserve population of a given First Nation is defined as the number of active band members living on a reserve or crown land that is associated with their own band according to the Indian Register on May 31st of a given census year.

### 4.1.2 Differences Between the Indian Register and Census Populations

The census population of a reserve is defined as the number of people living in a particular census subdivision (CSD) associated with a reserve as measured by the Census of Population in Canada for a given census year in the month of May. Statistics Canada defines a CSD as an “area that is a municipality or an area that is a deemed to be equivalent to a municipality for statistical reporting purposes (e.g., as an Indian reserve or an unorganized territory)”[17]. Ideally, the IR on-reserve population should approximate well the census reserve population if most reserves are inhabited mainly by members of the band that the reserves are associated with. The IR is updated more regularly than the census so it is expected to capture most of the changes in a reserve’s population since the previous census. However, there are many differences in population counts between these two sources.

If the IR on-reserve count is reflected perfectly by the census, then we can expect the census population of a reserve to be at least as large as the population according the IR. There could be people who are not listed on the IR that live on-reserve (e.g. members of other bands, people who have not registered with the IR yet, general population, etc.) that are captured by the census. There are also cases where people living on a reserve are not included in the IR. Often more than one factor contributes to differences across these data sources. Figure 4.1 shows
that the differences between the census and IR on-reserve populations for bands in 2011. Points that are below the blue identity line \((X = Y)\) are reserves where the IR population exceeds that of the census, suggesting that many people who are recorded as living on-reserve according to the IR may not be living there at the time of census collection.

Conversely, a number of reserves have populations according to the census that are much larger than the IR on-reserve population. The most extreme of these cases is the Westbank First Nation located in British Columbia which had 396 band members living on-reserve according to the IR in 2011, but 7068 people living there according to the 2011 Census [16]. Some First Nations such as the Westbank First Nation choose to lease parts of their reserve to non-band members and therefore could have larger populations relative to the number of registered band members living there [9][4]. Since the population according to the census is so much higher than the IR on-reserve population for these reserves, they represent possible outliers that may have high influence on the estimated parameters of regression models that we fit on the data.

**Figure 4.1:** 2011 Census Reserve Populations vs. 2011 IR On-Reserve Populations
There are many possible reasons why people listed as living on-reserve according to the IR may not appear in the census. The IR on-reserve population only considers registered band members who live on their own band’s reserve lands, but some reserves could be residences for members of other bands as well. The term “registered band members” refers to band members who are also registered on the IR. Some First Nations people may be registered as a member of a band and living on a reserve, but unless they are also listed on the IR, they are not included in the IR on-reserve population. The census captures the state of the country on the date of the census, which was May 16 in 2006 and May 10 in 2011 and 2016. However, the Indian Register data corresponds to May 31st of each census year. If changes occurred after census collection but before data was taken from the Indian Register, this would also lead to a discrepancy between the two counts.

In addition, a shortcoming of using the IR as an auxiliary data source is that it is usually updated via self-reporting by band offices [6]. These updates often occur after major life events of registered individuals such as births and deaths. A result of this is that the information contained can be outdated if bands or members do not self-report soon after these changes occur. In addition, many children are not registered with the IR as many of the benefits associated with being a Registered Indian are not relevant until a person reaches adulthood. This could lead to underestimation of census populations since children would be included in the census count if they were living on-reserve at the time. Deaths of Registered Indians are sometimes reported late or not at all since it is up to band or family members to notify INAC of these changes [7]. There are also possible issues related to the residence of Registered Indians as reported on the IR. Similar to other life changes such as births and deaths, moves on or off reserve may be reported late or not at all. If a person had a residence on-reserve and a residence off-reserve, it is possible that they are listed as on-reserve on the IR, but could be residing elsewhere on the date of the census [7]. All of these factors combined lead to discrepancies between the IR and census populations.

Although IR counts cannot be used to predict reserve populations directly, accurate predictions can still be made using an appropriate model.
4.2 Data Description

Data used for this case study was obtained from Statistics Canada and Indigenous and Northern Affairs Canada. INAC provided tables from the IR that listed each First Nation along with their own band on-reserve population for each census year from 2006 to 2016. Bands with an own band on-reserve population of less than 10 persons are censored due to data confidentiality concerns. The remainder of the data comes from tables that are publicly available on Statistics Canada’s website and through the software GeoSuite [13], also through Statistics Canada. The data obtained from Statistics Canada’s website includes census tables for 2011 and 2016 which contain the population count for each census subdivision for all three censuses of interest along with additional information such as the province of each subdivision and whether or not that area corresponds to a non-enumerated Indian reserve or not. This data was linked to the IR data using an additional table from Statistics Canada which lists all census subdivisions that are associated with each First Nation in Canada. Lastly, census metropolitan area (CMA) codes are obtained using GeoSuite. The CMA code is used to determine the census metropolitan influenced zone (MIZ) level of a reserve. Geosuite is a publicly available program that is provided by Statistics Canada that allows users to explore the links between all standard levels of geography and to identify geographic codes, names, unique identifiers, and, where applicable, types, as well as land area and population dwelling counts [14].

The response variable for each census year is the population of each reserve according to the census. The covariates include the IR on-reserve population as defined above, the province/territory that a band’s reserves are located in, and the weighted census MIZ level of a band’s reserves. The levels of provinces and territories include British Columbia (BC), Alberta (AB), Saskatchewan (SK), Manitoba (MB), Ontario (ON), Quebec (QC), Atlantic (AT), Yukon (YT) and the Northwest Territories (NT). The Atlantic provinces New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island each contain only a few reserves and are grouped together for this analysis. In addition, there are no First Nations located in Nunavut and therefore Nunavut is not listed as a level in our analysis under province/territory. It should be noted that according to the IR, the Taku River
Tlingit and Liard First Nation are both registered in the Yukon, however their associated reserves are located in BC according to the census. For this analysis, these bands will be considered as a part of BC.

The census MIZ variable is a measure of a reserves’ proximity to a metropolitan area measured on an ordinal scale from 0 to 5. Definitions of the census MIZ levels are shown in Table 4.1[17]. The reason for the inclusion of this variable is that bands with reserves located in or near census metropolitan areas/census agglomerations are hypothesized to have higher populations. CMAs are defined as areas consisting of one or more neighbouring municipalities situated around a core that has “a total population of at least 100,000 of which 50,000 or more live in the core”[17]. A census agglomeration (CA) is an area with “a core population of at least 10,000”[17]. Because some bands have multiple reserves that may be located in locations with differing proximity to a CMA, a rounded weighted average by the census population of each reserve in 2011 is used to compute the census MIZ level for a band. Since reserve locations do not change between censuses, the same values of province and census MIZ level are used for constructing regression models in 2006, 2011 and 2016. Although areas can potentially be reclassified as CMAs or CAs for the 2016 Census, it is unlikely to have any noticeable effect on our predictions.

<table>
<thead>
<tr>
<th>Level</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Located in a CMA or CA</td>
</tr>
<tr>
<td>1</td>
<td>Strong metropolitan influenced zone</td>
</tr>
<tr>
<td>2</td>
<td>Moderate metropolitan influenced zone</td>
</tr>
<tr>
<td>3</td>
<td>Weak metropolitan influenced zone</td>
</tr>
<tr>
<td>4</td>
<td>No metropolitan influenced zone</td>
</tr>
<tr>
<td>5</td>
<td>Located in a territory</td>
</tr>
</tbody>
</table>

Table 4.1: Levels of Census Metropolitan Influenced Zone

### 4.2.1 Out-of-Scope Bands

According to the data obtained from INAC, there are 633 unique First Nations in Canada as of May 2016. However, many of these bands and the reserves asso-
ciated with them are considered as out of scope for this thesis. Firstly, not every First Nation in Canada is associated with a reserve and therefore is not of interest to this analysis. As mentioned earlier, bands that had on-reserve populations of less than 10 are censored and these bands and their associated reserves are also not included in this analysis. In addition, some bands in the territories are associated with CSDs that are not considered Indian reserves. According to the census dictionary provided by Statistics Canada, the following CSD types are associated with on-reserve populations: Indian reserves (IRI), Indian settlements (S-É), Indian government districts (IGD), Terres réservées aux Cris (TC), Terres réservées aux Naskapis (TK), and Nisga’a land (NL) [17]. Many bands in the territories are associated with other CSD types, and they are considered out of scope. Lastly, some reserves in Canada are incompletely enumerated and do not have available counts for the 2006, 2011, or 2016 Censuses [19]. The bands associated with incompletely enumerated reserves are also considered out of scope for this analysis.

It should be noted that two reserves are associated with the Saddle Lake Cree Nation in Alberta, Saddle Lake 125 and White Fish Lake 128 [18]. White Fish Lake 128 is enumerated in 2006, 2011 and 2016, however Saddle Lake 125 is not [16] [15]. As a result, these reserves are also omitted from this study since the IR on-reserve population for Saddle Lake Cree Nation does not differentiate between the enumerated and non-enumerated reserve.

In the spring of 2011, The Narrows 49, the reserve associated with the First Nation Lake St. Martin in Manitoba experienced flooding [3]. According to the censuses in 2011 and 2016, the population of this reserve was 826 and 5, respectively. As the reserve is virtually uninhabited as of the 2016 Census, Lake St. Martin and its associated reserve are also excluded from this analysis.

4.2.2 Multi-Band Reserves and Multi-Reserve Bands

While most First Nations are associated with a single inhabited reserve, some First Nations are associated with multiple inhabited reserves. These bands are referred to as multi-reserve bands. Since the Indian Register does not differentiate which band members live on which reserve, it only states whether someone lives on a reserve that is associated with their own band, it is difficult to perform reserve spe-
cific predictions. Instead, populations of each reserve associated with a band are
totalled and modelling is done with the total reserve population for a band rather
than the individual reserve population. We make the assumption that each reserve
associated with that band grows proportionally between censuses. Predictions of
specific reserve populations are split based on the population ratios from the pre-
vious census. For example, suppose that a band with two reserves had a total of
100 people living on them according to the 2011 Census. If 70 people lived on
one reserve and 30 people lived on another, then the predicted total reserve popu-
lation for 2016 would be split at a ratio of 70 to 30 as well since we assume that all
reserves associated with a band grow at the same rate.

In addition, some reserves are associated with multiple bands, and are referred
to as multi-band reserves. Although the reserve only has a single population ac-
cording to the census, each band associated with a multi-band reserve has its own
on-reserve population according to the IR. To handle these reserves, each band that
shares a reserve has their IR population combined with other bands that share the
reserve and are treated as a single band when modelling and predicting. Since we
are only interested in reserve specific predictions, we do not need to split predic-
tions by band like we did for bands associated with multiple reserves.

A few reserves are associated with multiple bands which were also associated
with multiple reserves. These groups are referred to as multi-band multi-reserve
clusters in this thesis. An example of this are the Iskatewizaagegan #39 Indepen-
dent First Nation and the Shoal Lake No.40 First Nation, which are both associ-
ated with the Shoal Lake 34B2 reserve[18]. However, independently, they are also
associated with reserves that are exclusive to themselves. Iskatewizaagegan #39
Independent First Nation is associated with the reserves Shoal Lake (Part) 39A and
Shoal Lake (Part) 39A (Manitoba) while Shoal Lake No.40 First Nation is associ-
ated with the reserves Shoal Lake (Part) 40 and Shoal Lake (Part) 40 (Manitoba).
Since we do not know how many band members of each reserve reside on their
own reserve and on the shared reserve, we combine their IR on-reserve and census
populations together and treat these two bands as a single band. As with predic-
tions for multiple reserve bands, predictions for each individual reserve in 2016 are
split based on ratios observed for the 2011 Census.

Lastly, it should be noted that the Treaty Four Reserve Grounds 77 located
in the province of Saskatchewan are excluded from this study. This reserve is shared between 33 bands [8] and had a population of 10 people according to the 2011 Census [16]. If we considered this group of 33 bands sharing the reserve as a multi-band multi-reserve cluster, then our predictions would suffer from a loss of granularity. In this situation, omitting a reserve of around 10 people is less consequential compared to grouping 33 bands together and analyzing them as one cluster.

<table>
<thead>
<tr>
<th>First Nations</th>
<th>Associated Reserves</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iskatewizaagegan #39 Independent Shoal Lake No. 40</td>
<td>Shoal Lake (Part) 39A Shoal Lake (Part) 39A (Manitoba) Shoal Lake (Part) 40 Shoal Lake (Part) 40 (Manitoba) Shoal Lake 34B2</td>
</tr>
<tr>
<td>Bearsapw</td>
<td>Big Horn 144A</td>
</tr>
<tr>
<td>Chiniki</td>
<td>Eden Valley 216</td>
</tr>
<tr>
<td>Wesley</td>
<td>Stoney 142,143,144</td>
</tr>
<tr>
<td>Nisga’a Village of Gingolx Nisga’a Village of Gitwinksihlkw Nisga’a Village of Laxgalts’ap Nisga’a Village of New Aiyansh</td>
<td>Nisga’a</td>
</tr>
<tr>
<td>Aishihik Champagne</td>
<td>Champagne Landing 10 Kloo Lake Kluksu</td>
</tr>
</tbody>
</table>

**Table 4.2:** List of Multi-Band Reserves and Multi-Band Multi-Reserve Clusters

Table 4.2 lists the multi-band reserves and the multi-band multi-reserve clusters included in this case study. Non-enumerated bands in multi-band multi-reserve clusters such as the Mohawks of Kahnawà:ke and the Mohawks of Kanesatake, and the Six Nations of the Grand River considered out of scope and not listed in Table 4.2. In the list of Indian band areas and the census subdivisions they include [18], the Nisga’a Village of Gingolx, Nisga’a Village of Gitwinksihlkw, Nisga’a Village of Laxgalts’ap and Nisga’a Village of New Aiyansh are associated individually with the Nisga’a villages (NVL) Gingolx, Gitwinksihlkw, etc. However, accord-
ing to census dictionary in 2011, these villages make up the CSD, Nisga’a Lands (NL)\cite{17} and are treated as one reserve in this analysis. The Aishihik and Champagne First Nations are also associated with the CSD, Haines Junction. However, this CSD is considered a village (VL) \cite{18} and its population is not included in the census population for this multi-band multi-reserve cluster.

4.3 Prediction of Indian Reserve Populations in 2016

This section presents the prediction results of Indian reserve populations in Canada in 2016 using models compared in Chapter 3. Models are fit using data from the 2011 Census and IR and predictions are made using data from the 2016 IR. Census and IR data from 2006 are also used to check model assumptions and to fit a mixed effects model as well. Although one of the main advantages of temporal adjusted prediction over mixed effects models is that it does not require repeated measurements to estimate model parameters, since we have access to data for 2006, we compare the performance of a mixed effects model here too. Subsection 4.3.1 covers the exploratory analysis of the dataset and Subsection 4.3.2 presents the prediction performance results of each model in terms of MSPE. For this section, figures and numbers reported will be in terms of bands and not reserves unless otherwise specified. For example, Figure \ref{fig:4.1} compares the census and IR on-reserve populations of the 505 in-scope bands rather than the 771 reserves that they are associated with. This is done because the IR does not have reserve specific information for the bands that are associated with multiple reserves. However, since we are more interested in reserve populations instead of band populations, prediction results will be reported at the reserve level rather than at the band level.

4.3.1 Exploratory Analysis

After removing out-of-scope bands from the data, we are left with 505 bands which are associated 771 reserves. Table 4.3 shows a cross tabulation of the categorical variables province and census MIZ. BC and Ontario have the highest numbers of bands while the Yukon and Northwest Territories only include one band each. There are other bands located in the territories, but as mentioned in Subsection 4.2.1, they are out of scope since they are not associated with a CSD type that
corresponds to an Indian reserve. Most bands have reserves that are located in areas with weak or no metropolitan influence. BC has the highest proportion of bands that are located in either a census metropolitan area or a census agglomeration.

<table>
<thead>
<tr>
<th>Province</th>
<th>Census MIZ Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 1 2 3 4 5 Total</td>
</tr>
<tr>
<td>AB</td>
<td>7 0 14 5 14 0 40</td>
</tr>
<tr>
<td>AT</td>
<td>12 0 6 6 8 0 32</td>
</tr>
<tr>
<td>BC</td>
<td>49 3 16 24 90 0 182</td>
</tr>
<tr>
<td>MB</td>
<td>0 0 4 27 28 0 60</td>
</tr>
<tr>
<td>NT</td>
<td>0 0 0 0 0 1 1</td>
</tr>
<tr>
<td>ON</td>
<td>9 3 12 22 46 0 92</td>
</tr>
<tr>
<td>QC</td>
<td>4 0 2 18 7 0 31</td>
</tr>
<tr>
<td>SK</td>
<td>2 2 31 18 14 0 67</td>
</tr>
<tr>
<td>YT</td>
<td>0 0 0 0 0 1 1</td>
</tr>
<tr>
<td>Total</td>
<td>83 8 85 120 207 2 506</td>
</tr>
</tbody>
</table>

Table 4.3: Cross Tabulation of Province and Weighted Census MIZ

The first step of this analysis is to observe whether or not the assumption of a constant relationship between our covariates and response over time holds. If this condition isn’t reasonably satisfied, then our fitted models will likely not predict the population of Indian reserves in Canada well. When plotting the data in 2006 and 2011, shown in Figure 4.2, we notice a few things. Although our data follows a linear pattern in both 2006 and 2011, there are some potential outliers in the data that may have undue influence on the regression fit. This suggests that a robust regression model may be more appropriate for this dataset. In order to check whether or not the relationship between the covariates and response is constant over time, we fit a robust regression model on data from both years and compare the estimated coefficients shown in Table 4.4. The baseline levels of province and census MIZ are Alberta and 0 respectively. We are unable to estimate a coefficient for the level census MIZ of 5 since there is perfect collinearity between this variable and the levels of province, Yukon and Northwest Territories. Recall that a reserve with a census MIZ of 5 corresponds to a reserve located in a territory. In this case, we have only two bands located in the territories, one in the Yukon and the other
in the Northwest Territories. The estimated coefficients corresponding to these variables are able to capture the error relating to this level perfectly and therefore the coefficient for level of census MIZ of 5 cannot be estimated and is not included in Table 4.4.

Figure 4.2: Census Reserve Populations vs. IR On-Reserve Populations in 2006 and 2011

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\hat{\beta}_{i,2006}$</th>
<th>$\hat{\beta}_{i,2011}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>69.642</td>
<td>111.967</td>
</tr>
<tr>
<td>IR On-Reserve Pop.</td>
<td>0.841</td>
<td>0.814</td>
</tr>
<tr>
<td>Province:AT</td>
<td>33.289</td>
<td>-16.435</td>
</tr>
<tr>
<td>Province:BC</td>
<td>7.469</td>
<td>-22.928</td>
</tr>
<tr>
<td>Province:MB</td>
<td>-75.037</td>
<td>-88.959</td>
</tr>
<tr>
<td>Province:NT</td>
<td>-76.703</td>
<td>-29.765</td>
</tr>
<tr>
<td>Province:ON</td>
<td>3.461</td>
<td>-29.765</td>
</tr>
<tr>
<td>Province:QC</td>
<td>37.024</td>
<td>47.084</td>
</tr>
<tr>
<td>Province:SK</td>
<td>1.198</td>
<td>-24.773</td>
</tr>
<tr>
<td>Province:YT</td>
<td>-295.297</td>
<td>-147.485</td>
</tr>
<tr>
<td>Census MIZ:1</td>
<td>9.498</td>
<td>43.691</td>
</tr>
<tr>
<td>Census MIZ:2</td>
<td>-12.518</td>
<td>-22.674</td>
</tr>
<tr>
<td>Census MIZ:3</td>
<td>-17.447</td>
<td>-6.353</td>
</tr>
<tr>
<td>Census MIZ:4</td>
<td>-55.108</td>
<td>-60.631</td>
</tr>
</tbody>
</table>

Table 4.4: Estimated Robust Regression Coefficients for 2006 and 2011 Reserve Populations

The estimated coefficients related to census MIZ are similar in both 2006 and 2011 across all levels. The estimated intercept for each year is the estimated reserve
population according to the census for a band with an IR on-reserve population of 0, located in the province of Alberta and within a CA or CMA. The estimated coefficient for the IR on-reserve population indicates the expected increase in the census population of a reserve for a one person increase in the IR on-reserve population for the corresponding year, while holding all other variables constant. The estimated coefficients for province and census MIZ are the estimated effects of the levels of province and census MIZ relative to the baseline. For example, the coefficient corresponding to Ontario in 2011 is -29.765, meaning that a band with reserves located in Ontario is expected to have 29.765 less people on the census in 2011 compared to a band in Alberta who share the same values of its other covariates. One interesting thing to note is that reserves located near urban areas, or lower levels of census MIZ, correspond to higher populations while more reserves in more remote areas correspond to lower populations. The most important difference between the two fitted models is the difference between the estimated coefficients for the IR on-reserve populations in 2006 and 2011 because this covariate is highly correlated with the response. The difference in the estimated coefficients between 2006 and 2011 is 0.027. As shown previously in Section 3.2.4, this could have implications on the prediction accuracy of each model if the estimates in 2006 and 2011 do not accurately reflect the relationship between census and IR on-reserve populations in 2016. If the true regression parameters in 2016 are similar to the estimates in 2011, then we can expect predictions of reserve populations for 2016 from temporal adjusted prediction to be accurate. Although this is not a formal test of a constant relationship between the covariates and response over time, it is reasonable to expect that the model will not change drastically in 2016 as well.

4.3.2 Prediction Results

We fit each of the models compared in Chapter 3 on the 2011 Census and IR data and then predict the 2016 Census population of each reserve using the fitted model and the IR on-reserve populations from 2016. We also include the predictions results of the LME model which was fit using both 2006 and 2011 data since it requires repeated measures. Figure 4.3 shows boxplots comparing the prediction
performance of each model, and Table 4.5 shows the MSPE of each model when predicting the populations of each Indian reserve. Note that the MSPE reported in Table 4.5 differs from the MSPE reported in tables in Chapter 3. Tables in Chapter 3 reported the median MSPE of 1000 simulated datasets, while Table 4.5 reports the MSPE of a single dataset.

**Figure 4.3:** Comparison of Prediction Error of 2016 Reserve Populations

Figure 4.4 shows a truncated view of Figure 4.3 so that we can see the distribution of prediction errors more closely. Most of the prediction errors for each method seen in Figure 4.4 are close to zero, although linear regression and robust regression appear to have the largest spread. These results are consistent with the simulation results shown in Chapter 3 where methods like TAP, robust TAP, linear extrapolation and LME are able to predict better than linear regression.

Linear extrapolation was able to predict with the highest accuracy among all of the selected models in terms of MSPE, followed by robust TAP and TAP, then LME, and lastly linear regression and robust regression. It is interesting to see that linear extrapolation predicts the best in terms of MSPE when it consistently
Figure 4.4: Comparison of Prediction Error of 2016 Reserve Populations (Zoomed In)

predicts worse than TAP and robust TAP in our simulations. On closer inspection, this result is likely due the fact that extrapolation is able to predict the population of the outliers better than the other models. As shown in Figure 4.3, each model had a large negative residual which corresponds to the Tsinstikeptum 9 reserve in BC. The Tsinstikeptum 9 and Tsinstikeptum 10 reserves are both associated with the Westbank First Nation which was mentioned earlier in this chapter as an outlier.
where the total population of its reserves as measured by the census greatly exceeds
the IR on-reserve population for each census year. Although none of the models are
able to predict the population of Tsinstikeptum 9 well, linear extrapolation comes
the closest. The residuals for Tsinstikeptum 9 are shown in Figure 4.3 as the lowest
point in each boxplot. The poor prediction of just one reserve can considerably
inflate the MSPE as shown by each model in Table 4.5.

Since there are outliers present in the data, we also compare another measure
of prediction error that isn’t as severely affected by large residuals, the median
absolute prediction error (MAPE). This measure is calculated in a similar manner
to MSPE, except we take the median of the absolute value of the residuals rather
than the mean of the squared residuals. These values are also shown in Table 4.5.

Robust TAP predicts with the lowest MAPE followed closely by TAP, LME,
linear extrapolation and then by robust regression and linear regression. These
results do not exactly align with those observed in any of the scenarios discussed
in Subsection 3.2.4. When fitting a robust regression model on the 2016 data as
we did on the 2006 and 2011 data in Subsection 4.3.1 the estimated coefficient for
the IR on-reserve population is 0.829. This means that our dataset most closely
resembles Scenario (3) of Subsection 3.2.4 where the true coefficient $\beta_1$ is not
constant over time, although the results from the simulation and case study differ.
In that simulation, it was shown that the prediction accuracy of all models are
affected because the true regression coefficient was different in the third time point
compared to the first two time points. However, LME is the most badly affected
model among TAP, robust TAP and extrapolation. In this case study, we see that

<table>
<thead>
<tr>
<th>Method</th>
<th>MSPE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extrapolation</td>
<td>14110.14</td>
<td>21.79</td>
</tr>
<tr>
<td>Linear Mixed Effects Model</td>
<td>17425.72</td>
<td>21.37</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>128787.89</td>
<td>58.45</td>
</tr>
<tr>
<td>Robust Regression</td>
<td>139002.38</td>
<td>36.92</td>
</tr>
<tr>
<td>Robust Temporal Adjusted Prediction</td>
<td>15617.08</td>
<td>20.30</td>
</tr>
<tr>
<td>Temporal Adjusted Prediction</td>
<td>15682.89</td>
<td>20.47</td>
</tr>
</tbody>
</table>

Table 4.5: Mean Square and Median Absolute Prediction Errors of 2016 In-
dian Reserve Populations
LME is not as badly affected as it was in the simulations. However, under both performance metrics, LME still predicts worse than TAP and robust TAP in this case study.

Although we can look at different performance metrics that show one model is better than another, from the results of this case study, we can at least conclude that the prediction accuracy of TAP is comparable to that of the LME model. We show that TAP is able to produce similar prediction results as LME with the improvement that it only requires data from a single time point to fit the model.
Chapter 5

Conclusion

In this thesis we propose a new prediction model, temporal adjusted prediction, which is suitable for prediction of responses in datasets with observations for a set of individuals at multiple time points. Although linear mixed effects models can be used in some instances, it requires repeated measures in order to fit a model. In situations when repeated measures are not available, other models such as linear regression can be fit using data from a single time point. However, they do not take into account the within-individual correlation when making predictions. In cases where observations are made on the same set of individuals, future values of a response will be correlated with past values of the response.

As shown in Section 2.4, predictions using TAP are based on a random intercept model even though it does not directly predict the random intercept term. Instead, the previous response values are used as starting points for predictions. A consequence of this is that additional error is introduced into the predictions. However, if the variance of the random intercept is greater than the variance of the random error, then the expected MSPE of TAP is lower than that of linear regression. As shown in simulations in Chapter 3 and the case study in Chapter 4, even when repeated measures are available, TAP is shown to have similar prediction accuracy compared to LME models with parameters estimated using maximum likelihood methods despite requiring less data to fit a model.

There are many possible extensions of the work presented in this thesis. Further simulations can be done comparing TAP to other models, especially linear mixed
effect models, and other performance metrics can be examined as well. When looking at MSPE alone, it was not conclusive that TAP or LME models were superior to each other. Next, further research can be done on robust TAP by using different estimation methods such as M-estimation, S-estimation, or least trimmed squares to estimate the regression parameters. In addition, the simulations in this thesis only examined the effects of outliers in the response because these were similar to the outliers observed in the Indian reserves population data. However, research can be done on how extreme values in the covariates affect the overall fit and prediction accuracy of robust TAP. Finally, one of the main weaknesses of TAP is that it does not predict as well as linear regression when the variation due to the random intercept is small compared to random error. Improvements in this situation will make TAP more widely applicable.
Bibliography


Table-Tableau.cfm?LANG=Eng&T=301&S=3&O=D. Last updated August 9, 2016. (accessed May 25, 2017). → pages 2, 41, 44, 48, 50


