Travel Distance and Land Use: Application of a Generalized Box-Cox Model with Conditional Spatial Lag Dependence

Jason Hawkins *[1] and Khandker Nurul Habib[1]

[1]Department of Civil Engineering, University of Toronto, Toronto, ON

January 4, 2019

Abstract

The relationship between travel distance and land use factors has seen extensive study in recent years. It is clear that land use policy influences the distance we have to travel to reach the locations of our daily activities. In this study we focus on applying several econometric techniques, some of which have been applied in isolation in past studies, but the combination of which have not been applied to the problem. We develop a series of statistical tests using the method of artificial regression to test the joint effects of functional form and spatial autocorrelation on model fit. Application is made to the Greater Toronto Area via a large scale travel survey administered in 2016. Several unexpected results are obtained, namely that there is no spatial dependence in travel distances for this region and that population density tends to increase the distance travelled by respondents. Model results are validated using bootstrapped sampling and local indicators of spatial association.

1 Introduction

One of the principal objectives of land use policy at the metropolitan scale is to address its interaction with the transportation network. Land use influences transportation behaviour through the modes of travel and diversity of activities available in a given area. It also has an indirect effect through its influence on trip lengths, which are defined by the distances between daily activities (e.g. home, work, shopping). Disentangling the effects of land use and household factors on daily travel distance is an ongoing challenge in the transportation analysis field [12, 14, 20, 23]. In addition to defining model variables, the analyst is faced with a variety of econometric and computational challenges in the development of a suitable model.

One question, which has garnered little attention, is the functional form of the relationship between exogenous variables and the endogenous variable of travel distance. A simplifying assumption is that the relationship is linear but, as in the case of many socioeconomic systems, it is likely that there is a diminishing marginal effect. For example, there is ample evidence that increasing household income is associated with additional travel [15, 21], but it is likely that the marginal utility is sublinear. Many studies address this effect through an *a priori* logarithmic relationship, but do not justify this assumption through statistical tests. A second common adjustment in such models is for potential spatial autocorrelation between observations. It is a reasonable hypothesis that proximate households will, *ceterus paribus*, have similar transportation patterns. This can be considered through the introduction of a spatially correlated error term or a spatially lagged dependent variable. However, the inclusion of both these econometric instruments has not be tested, to the best of our knowledge. The introduction of spatial lag dependence requires the definition of a square spatial weight matrix with dimensions defined by the number of observations. This introduces computational challenges as the manipulation, particularly inversion, of large matrices requires the storage of large amounts of data in active memory. Some authors have developed software using sparse matrix representations to address this challenge [3, 17], but their functionality is limited and they do not address challenges in initial

*Corresponding author

data preparation for large datasets. We propose bootstrapping as a means of maintaining the advantages of large datasets, while reducing the computational burden associated with each matrix operation.

In the present study, we seek to combine all the above noted methods to produce a robust econometric model of travel distance as a function of land use and household characteristics. We build upon initial work by [10, 6] to develop statistical tests for the combined application of a Box-Cox transformation to account for functional form and lagged dependence to account for spatial autocorrelation. The model is applied to a large scale travel survey for the Greater Toronto Area (GTA), representing roughly 162,000 household records from the 2016 Transportation Tomorrow Survey (TTS). We apply the econometric model to a random sample of 2000 respondents and use bootstrapping to prove the robustness of the results for the full TTS dataset. Extensions to the spatial lag dependence are made through an interaction with household demographic variables. This produces a spatial weight matrix, wherein the weights are conditional upon the characteristics of the adjacent household.

2 Literature Review

A particularly pertinent study is by Morency et al. [23] who examine the question of travel distance in the same region (i.e. Greater Toronto Region) using 2001 TTS data. They use a spatial expansion method, based on work by Casetti [7], which separates variables into common and spatially varying components. Spatial variation is accomplished by interacting the variable of interest with the longitude and latitude of the origin location. Morency et al. [23] apply a log transformation to travel distance and perform their analysis to all trips in the TTS. The spatial expansion method allows for the calculation of a parameter surface for the study region, whereby the parameter value may vary across each spatial dimension. Their analysis focuses on travel distance by the elderly, individuals in single-parent households, and low-income households. They find that travel distances tend to be lowest among residents of the central city, and increase towards the north-east of the Greater Toronto Area. A shortcoming of the study is that spatial statistical method employed only captures *own* spatial effects and not the influence of social effect of spatially proximate households.

Kasraian et al. [16] examine the evolution of travel distance in the Dutch context using a 30 year pseudo-panel. The geographic scope of this study is larger, a small European nation, so that trip origins and destinations are reported at the level of municipalities and 4-digit postal codes for records from the years 2005 and 2010. They estimate the model using 21 representative groups for each survey year and a hybrid OLS regression design. Independent variables are demeaned within each survey year against the respective group mean and estimated using random effects. The dependent variable is estimated as the untransformed travel distance for each survey respondent.

De Abreu Silva et al. [11] also apply a log transformation to commuting distance in their analysis of transportation patterns in Montreal. Their work represents one of several studies of travel distance in Montreal that uses structural equation models [11, 19, 22]. Manaugh et al. [19] develop a set of neighbourhood classification variables using factor and cluster algorithms. They classify 150m squared grid cells using a large set of land use variables. This can be roughly termed as a structural equation model because the factor-cluster variables are endogenous variables of the base land use variables, and subsequently used in the estimation of the travel distance model. Miranda-Moreno et al. [22] use a similar set of techniques to simultaneously model car ownership and neighbourhood typology as inputs to a joint residential location and auto ownership choice, which then influences the choice of average travel distance.

Ellder [13] focuses on the question of travel distance differentiated by trip purpose and how each is influenced by residential location choice. He uses a hierarchical OLS model, which employs individual-specific and neighbourhood-specific error terms. This has the effect of producing a random intercept term in the regression, which varies by neighbourhood. A shortcoming of this model is that it is restricted by the MAUP in its definition of each neighbourhood.

Aditjandra et al. [1] examine the effect of neighbourhood characteristics on vehicle km driven by residents of north-east England. They use the untransformed average trip distance in a weekly travel survey.

Ding et al. [12] use a hazard regression model to consider the influence of built environment characteristics on commuting distance. The trouble with this approach is that it is also, typically, bounded by *a prior* assumptions about the distribution of commuting distance on the part of the analyst. The hazard function is often parameterized as a Gumbel or Gopertz distribution, but extension could be made to a generalized

gamma distribution, which would act similar to the Box-Cox model to provide a flexible parameterization of the dependent variable. A second note on the hazard regression method is that its origins are in health science applications where it is common to not observe the beginning and/or termination of the dependent variable (i.e. left and right censoring, respectively). This does not necessarily limit its application to travel distance regressions, but the travel distance is always observed in such analysis and therefore the hazard function only provides a means of variable transformation.

3 Data Sources

The main data source for this study is the Transportation Tommorrow Survey (TTS), a household travel survey conducted every 5 years for residents of the Greater Golden Horseshoe Region of Ontario (including the GTA). The TTS is a 5% sample of the regional population and is among the largest and longest running travel surveys. Figure 1 shows the five regions contained within the study area.

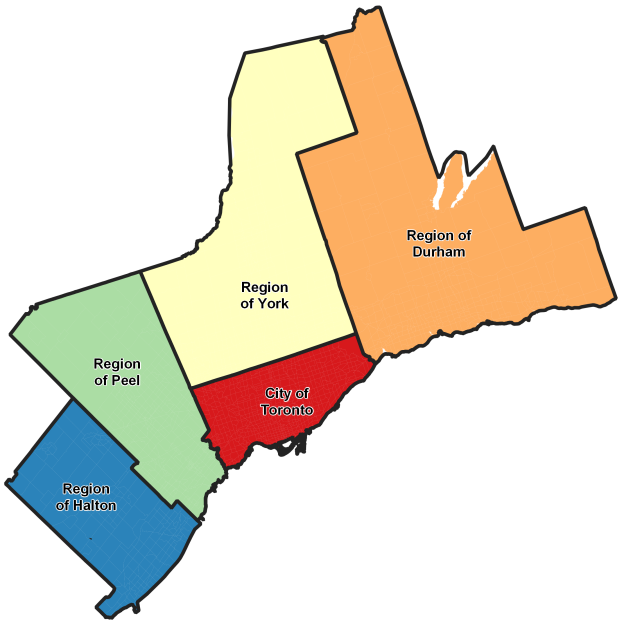

Figure 1: TTS regions included in analysis.

A secondary data source from the Toronto Real Estate Board (TREB) provides average dwelling price data for each TAZ in the study area.

4 Methods

The study of the relationship between travel distance and land use is not new ground [4, 11, 13, 14, 20, 22, 23]. We examine this question as one of continued interest in the fields of transportation and land use analysis, but place an equal focus on the application of various econometric techniques to the problem.

The first technique we explore is spatial lag dependence on the dependent variable, distance. It is a well supported hypothesis that travel distances are related for households located in the same community. Such households will be similarly constrained by their proximity to retail and potential work locations. This dependence manifests in regression models of travel distance as an autocorrelation between observations. The inclusion of a spatial lag term allows the analyst to test the interaction between travel distance of spatially proximate households. The inclusion of a spatial lag term is relatively straightforward and existing software exists in Python, Matlab, R, among other packages. A standard representation of a spatial lag model is

$$\boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where ρ is a measure of spatial correlation, $\boldsymbol{W}$ is a spatial weights matrix, $\boldsymbol{X}$ is a matrix of exogenous variables and ϵ is a non-spatially varying error term.

The above formulation assumes that observations are similarly influenced by all adjacent observations. However, it is often the case that additional information is available about the adjacent observations (i.e. households) that influences the spatial autocorrelation. For example, households with similar incomes may exhibit a stronger spatial correlation in their travel distances. We explore the inclusion of an interaction term between sociodemographic explanatory variables and the spatial weights matrix. This method offers a simple means of interacting spatial and sociodemographic attributes of each household, but reduces the number of spatially weighted terms with the application of each additional sociodemographic variable. As such, we only consider a single interaction (e.g. interacting the spatial weights matrix with the adjacent households having similar incomes). A more complex formulation would parameterized ρ as a function of a series of sociodemographic variables. This introduces additional challenges in model estimation due to the nature of the spatial lag model. The reduced form spatial lag model is

$$\boldsymbol{y} = (\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{\epsilon} \tag{2}$$

which means that $\boldsymbol{W}\boldsymbol{y}$ contains $\boldsymbol{W}(\boldsymbol{I}-\boldsymbol{\rho}\boldsymbol{W})^{-1}\epsilon$ and is therefore correlated with ϵ. The loglikelihood function is therefore

$$LL = -\frac{n}{2}\ln(2\pi) - (1/2)\ln(\sigma^2) + \ln|\boldsymbol{I} - \rho\boldsymbol{W}| - (\boldsymbol{y} - \rho\boldsymbol{W}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \rho\boldsymbol{W}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/2\sigma^2 \tag{3}$$

assuming $\epsilon \sim N(0, \sigma)$. The Jacobian transformation becomes part of the loglikelihood function and optimization can not be performed as the summation of loglikelihoods. Anselin [2] defines the problem as a single variable non-linear optimization in ρ by deriving values for β and σ, then substituting these values into a concentrated loglikelihood function to obtain

$$LL = \ln|\boldsymbol{I} - \rho\boldsymbol{W}| - \frac{n}{2}\ln[(\epsilon_0 - \rho\epsilon_L)'(\epsilon_0 - \rho\epsilon_L)/n] \tag{4}$$

where

$$\boldsymbol{\epsilon_0} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \tag{5}$$

$$\boldsymbol{\epsilon_L} = \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'W}\boldsymbol{y} \tag{6}$$

As the loglikelihood includes a Jacobian term, the information matrix cannot be determined using the normal simplifications and partial derivatives must be determined explicitly. With these conditions, it is clear that parameterizing the spatial lag is not a simple process.

The second component of the model is the explicit estimation of its function form via Box-Cox transformation of the dependent and continuous explanatory variables. Previous models have assumed an *a priori* linear or loglinear functional form, but is typically the case that this assumption is not tested. The functional form provides valuable information as to the nature of the influence of explanatory variables on travel distance. Estimation of the Box-Cox parameter provides the analyst with a robust measure of the growing/decaying influence of explanatory variables on travel distance. The resulting combination of spatial lag dependence and Box-Cox transformation produces a complex model, which can be estimated as a two-parameter loglikelihood function. Baltagi and Li [5] derive the hessian matrix for the model, from which the information matrix can be approximated as its negative.

Specification tests for the model are provided by [5] using the Lagrange multiplier method. These tests are cumbersome to program and modify for different model specifications. Fortunately, Li and Le [18] derive a simplified version of the tests using the method of double-length artificial regressions. For the details of the asymptotic properties of this test, refer to [9] and [18]. The test employs an artificial regression of the restricted model with a vector of ones appended to the dependent variable vector. The loglikelihood function is decomposed into two components: $\boldsymbol{f}$ representing the standardized error vector for the OLS regression model and $\boldsymbol{k}$ representing the additional terms from the Jacobian transformation. In case of the generalized Box-Cox model, with spatial lag dependence, the loglikelihood function is

$$LL = -\frac{n}{2}\ln 2\pi + \sum_{i=1}^{n} k_i(y,\theta) + \frac{1}{2}\sum_{i=1}^{n} f_i^2(y,\theta) \tag{7}$$

where

$$f(y,\theta) = \frac{1}{\theta}[(I - \rho W)y^{(\lambda)} - X^{(\lambda)}\beta - Z\gamma] \tag{8}$$

$$k(y,\theta) = -\ln\sigma + \ln(1 - \rho\omega_i) + (\lambda - 1)\ln y_i \tag{9}$$

In the above equations, θ is the set of all parameters, ρ is the spatial lag term, λ is the Box-Cox term, ω_i is the eigenvalue for the weights associated with observation i, and $y^{(\lambda)}$ is (similar for $X^{(\lambda)}$)

$$y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda} \tag{10}$$

From this, we can define sets of partial derivatives $\boldsymbol{F}(y,\theta)$ and $\boldsymbol{K}(y,\theta)$

$$\boldsymbol{F} = [-\frac{1}{\sigma^2}((\boldsymbol{I} - \rho\boldsymbol{W})\boldsymbol{y}^{(\lambda)} - \boldsymbol{X}^{(lambda)}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\gamma}), -\frac{1}{\sigma}\boldsymbol{X}^{(\lambda)}, -1\frac{1}{\sigma}\boldsymbol{Z}, -\frac{1}{\sigma}\boldsymbol{W}\boldsymbol{y}^{(\lambda)}, \frac{1}{\sigma}((\boldsymbol{I} - \rho\boldsymbol{W})C(y,\lambda) - C(X,\lambda)\beta] \tag{11}$$

where $C(y,\lambda) = \frac{1}{\lambda^2}(\lambda\boldsymbol{y}^\lambda \odot \ln y - \boldsymbol{y}^\lambda + 1)$ and $\odot$ denotes the Hadamard product. Note that $C(y,0) = \lim_{\lambda->0} C(y,\lambda) = \frac{1}{2}(\ln y)^2$ and $C(y,1) = y \odot \ln y - y + 1$. The ith row of $K(y,\theta)$ is

$$\boldsymbol{K} = [\frac{1}{\sigma}, \boldsymbol{0}_{1xK}, \boldsymbol{0}_{1xL}, -\frac{\omega_i}{1 - \rho\omega_i}, \ln y_i] \tag{12}$$

where $\boldsymbol{0}_{1xK}$ and $\boldsymbol{0}_{1xL}$ are vectors of zeros for the K Box-Cox transformed exogenous variables and L non-Box-Cox transformed exogenous variables, respectively. These results are essentially first-order order derivatives of components of the loglikelihood function and the double length regression method involves solving the following equation

$$\begin{bmatrix} f(y,\theta) \\ i_n \end{bmatrix} = \begin{bmatrix} -\boldsymbol{F}(y,\theta) \\ \boldsymbol{K}(y,\theta) \end{bmatrix} \delta + \epsilon \tag{13}$$

Using several properties of the loglikelihood function, we obtain an LM test of the following form

$$(-\hat{f}'\hat{\boldsymbol{F}} + i_n'\hat{\boldsymbol{K}})(\boldsymbol{F}'\boldsymbol{F} + \boldsymbol{K}'K)^{-1}(-\boldsymbol{F}'f + \boldsymbol{K}'i_n) \tag{14}$$

LM tests can be derived from the above equation by estimating restricted versions of the base model. The test statistics is asymptotically distributed as χ^2_R where R is the number of restrictions in H_0.

We consider a set of 9 tests, with DLR parameters as given in [18]. The H_0 for each case are outlined in table 4:

Table 1: DLR Test Statistics

Test Hypothesis	Description
	Joint Tests
H_0^1: $\rho = 0$ and $\lambda = 0$	Loglinear model without spatial lag dependence against a generalized Box-Cox model with spatial lag dependence.
H_0^2: $\rho = 0$ and $\lambda = 1$	Linear model without spatial lag dependence against a generalized Box-Cox model with spatial lag dependence.
	One-Directional Tests
H_0^3: $\rho = 0 \vert \lambda = 0$	Without spatial lag dependence, assuming a loglinear model.
H_0^4: $\rho = 0 \vert \lambda = 1$	Without spatial lag dependence, assuming a linear model.
H_0^5: $\lambda = 0 \vert \rho = 0$	Loglinear model against a generalized Box-Cox model, assuming no spatial lag dependence.
H_0^6: $\lambda = 1 \vert \rho = 0$	Linear model against a generalized Box-Cox model, assuming no spatial lag dependence.
	Conditional Tests
H_0^7: $\rho = 0$ — λ unknown	No spatial lag dependence, assuming a Box-Cox model.
H_0^8: $\lambda = 0$ and ρ unknown	Loglinear model against a generalized Box-Cox model, with possible spatial lag dependence.
H_0^9: $\lambda = 1$ and ρ unknown	Linear model against a generalized Box-Cox model, with possible spatial lag dependence.

In the above listing, we differentiate joint tests as an explicit consideration of the parameter value, while the one-directional test is an implicit exclusion by the analyst (typical of most applications). For example, we can test for an absence of spatial lag dependence in the data and a linear structure as in test #2. In contrast, test #4 considers the case that spatial lag dependence is tested for the model, but it is assumed *a priori* that the function form is linear.

The final method explored in this research is bootstrap sampling. This is a powerful method employed in computer science and econometric, which has seen limited application in transportation demand analysis. In this case, we use bootstrapping to overcome computational limitations, while fully utilizing the available dataset. We use approximately 162,000 household records from the 2016 TTS survey. The spatial lag model requires a weights matrix, which has dimensions of nXn where n is the number of observations. This introduces a wide array of data preparation and model estimation challenges including: conversion to sparse matrices, upwards of 26 trillion distance calculations between observation pairs, and inversion of a matrix that cannot be easily stored in RAM. A standard method is to take a sample of observations and perform analysis on these data. However, an objective of this research was to fully utilize the complete set of records. Bootstrapping provides a means of bypassing the need to estimate a model with such a large weight matrix. We first limited the number of neighbours included for each observation to 300 (maintaining 48.6 million non-zero data points in the weight matrix). A total of 100 random samples, each containing 2000 households observations, were drawn from the TTS survey and weight matrices developed for each sample (requiring a smaller matrix of 2000 x 2000 for each). This method allows us to estimate each model in a reasonable time (approximately 5 seconds per estimation) and devise confidence intervals for parameter values and standard errors for each of the estimated parameters. Even with the computational capacity to perform the larger matrix inversion, most of the operations are not O(n=162,000) and therefore the estimation process would be considerably faster (i.e. 100 x O(n=2000^k), where k is a function of n for each operation).

5 Results

An initial set of regression models were estimated, which varied the assumptions made about the spatial dependence term. The first dimension of this variation was whether to use the inverse distance or a threshold distance of 1 km. The second dimension was conditioning of the spatial weight matrix on sociodemographic variables. Comparisons were based on the AIC measure and it was determined that an inverse distance matrix does not offer improvements over a simple 1 km distance threshold. With respect to conditioning of the spatial weight matrix on sociodemographic variables, dwelling type, the presence of children in the household, and household income were explored for each of the three dependent variables. The AIC measures indicated that only income offered an improvement over an unconditional spatial weight matrix, and only for total travel distance. The specification with the lowest AIC measure is used in subsequent analysis for each of the three models. For each model, only non-zero records are considered in the estimation.

With the combination of a Box-Cox transformation and spatial lag dependence, none of the models exhibit significant spatial dependence. We explored weight matrix formulations based on a distance band of 1 km and inverse distance. Part of the reason for this lack of significant spatial autocorrelation may be that distances are only calculated for the 300 closest households for each record. By drawing 2000 records from the full dataset, this is further reduced in each estimated model. However, we find that there are still an average of 6 non-zero entries for each record. In most instances, 300 records encompasses all records in the same TAZ; however, there are a few dense zones with upwards of 450 records in the same TAZ.

For each of the three models we will discuss the spatial lag dependence and Box-Cox functional form, LM test statistics, and consider individual parameter values. Following this discussion of estimation results, bootstrap plots will be presented and discussed.

5.1 Total travel distance

Total travel distance exhibits a functional form that is clearly sub-linear with respect to the continuous independently variables, but also statistically different from a natural logarithmic transformation. There is no evidence of spatial dependence in the untransformed travel distance of survey respondents. However, the LM test statistics warrant the inclusion of spatial dependence in the model specification when applied in

combination with a Box-Cox transformation of travel distance and continuous independent variables. It is also evident that a Box-Cox transformation is warranted, regardless of the existence of spatial dependence.

Average dwelling price tends to decrease total travel distance, which suggests that residents of the GTA are willing to pay a premium to reduce their travel distance. The positive sign for population density fits with previous findings by Morency et al. [23] for Toronto. They find the opposite sign for similar models in Hamilton and Montreal, and suggest that the lower population densities in these regions is a potential cause for the difference. The high population density of Toronto is likely associated with higher rates of non-motorized travel, which tends to be under-reported in household travel surveys [23]. This may be a partial cause for the positive sign of this variable, but our estimation only includes the principal respondent (i.e. removes surrogate bias) and the effect of non-motorized travel in higher density areas is unlikely to fully explain this result. A second cause may be the high price of real estate in the denser communities of Toronto pushing out commercial activity. There is a high density of work and residential development within central Toronto, but a lack of large appliance and furniture stores. Residents of these higher density areas may face longer travel distances to these shopping opportunities.

The model suggests that males tend to travel further than females, which is supported by past research findings of longer commute distances for males [8]. In support of the hypothesis that under-reporting of non-motorized trips tends to bias the sign of population density, active trip count has a negative and significant parameter. This suggests that active trips, more common among residents of higher density areas, tend to be shorter than motorized trips (as one would expect). Interestingly, an additional vehicle seems to have the same effect on travel distance as the possession of a GO or PRESTO transit pass. In the case of additional vehicles, this indicates a household that is auto dependent and likely located in a low density area where average trip distances are longer. Possession of a GO pass suggests a long commute from a suburban community into the Toronto CBD. These two variables may be associated with the same households, that is auto dependent households who possess multiple vehicles also tend to commute by GO train. Local transit users do not tend to have longer travel distance than otherwise similar persons. As one would expect, non-workers tend to travel less because they do not need to make a commuting trip.

Despite the model indicating weak spatial dependence, this does not mean that there are no spatial effects on travel distance. Global differences in population and employment density, transit access, and spatial configuration between regions were also explored through categorical variables for each region. The City of Toronto was taken as the reference region as it is the most central and the nucleus of employment for the GTA. This provides an interesting set of results, which fits with the spatial configuration of the GTA. Residents of Durham and Peel have longer travel distances than other regions. Figure 1 shows that these regions lie adjacent to the City of Toronto and are long strips extending from their southern border along the Lake Ontario shoreline. This shoreline is the focus of urban development and the major GO Train services. As such, residents of these two regions must generally travel south to reach their places of employment. To the north, the region of York has urban centers more evenly distributed throughout and strong transit connection through the TTC subway into the City of Toronto. To the east, the Region of Halton of less reliant on the City of Toronto because it is closer to the City of Hamilton, thus reducing the average trip distance of respondents in this region.

Examining Figure 2, spatial dependence does not appear to be a factor in any of the 100 bootstrapped samples. The Box-Cox transformation shows minimal variation between samples, which suggests the functional form determined for the estimation sample fits the overall TTS dataset. Most of the parameters show little variation between samples and estimation results seem to be quite strong.

Table 2: Summary of results for total travel distance

Variable	Param	Std Err.	t-stat	p-value
Constant	0.06	0.074	0.754	0.451
Average dwelling price in TAZ	-0.14	0.068	-2.062	0.039
Population density of TAZ (persons per km^2)	4.02	0.485	8.29	0.000
Gender	0.21	0.085	2.465	0.014
Active trip count	-0.06	0.012	-5.009	0.000
Transit pass (GO/PRESTO)	0.77	0.118	6.549	0.000
Transit pass (local)	-0.04	0.168	-0.265	0.791
Vehicles per driver	0.74	0.108	6.807	0.000
No employment	-0.83	0.094	-8.782	0.000
Home in Durham or Peel region	0.35	0.150	2.326	0.020
ρ	0.01	0.012	0.800	0.424
λ	0.16	0.005	32.749	0.000
Log likelihood		-7643.3		
AIC		15310.7		
Number of observations		1669		
Pseudo R-squared		0.088		

LM Test Statistics

H_0	χ^2	p-value
Joint $\rho = 0$ and $\lambda = 0$	444.6	0
Joint $\rho = 0$ and $\lambda = 1$	1174978	0
1D $\rho = 0 \vert \lambda = 0$	444.5	0
1D $\rho = 0 \vert \lambda = 1$	1174978	0
1D $\lambda = 0 \vert \rho = 0$	444.2	0
1D $\lambda = 1 \vert \rho = 0$	1174978	0
Cond. $\rho = 0$ and λ unknown	1.1	0.304
Cond. ρ unknown and $\lambda = 0$	445.9	0
Cond. ρ unknown and $\lambda = 1$	1175332.1	0

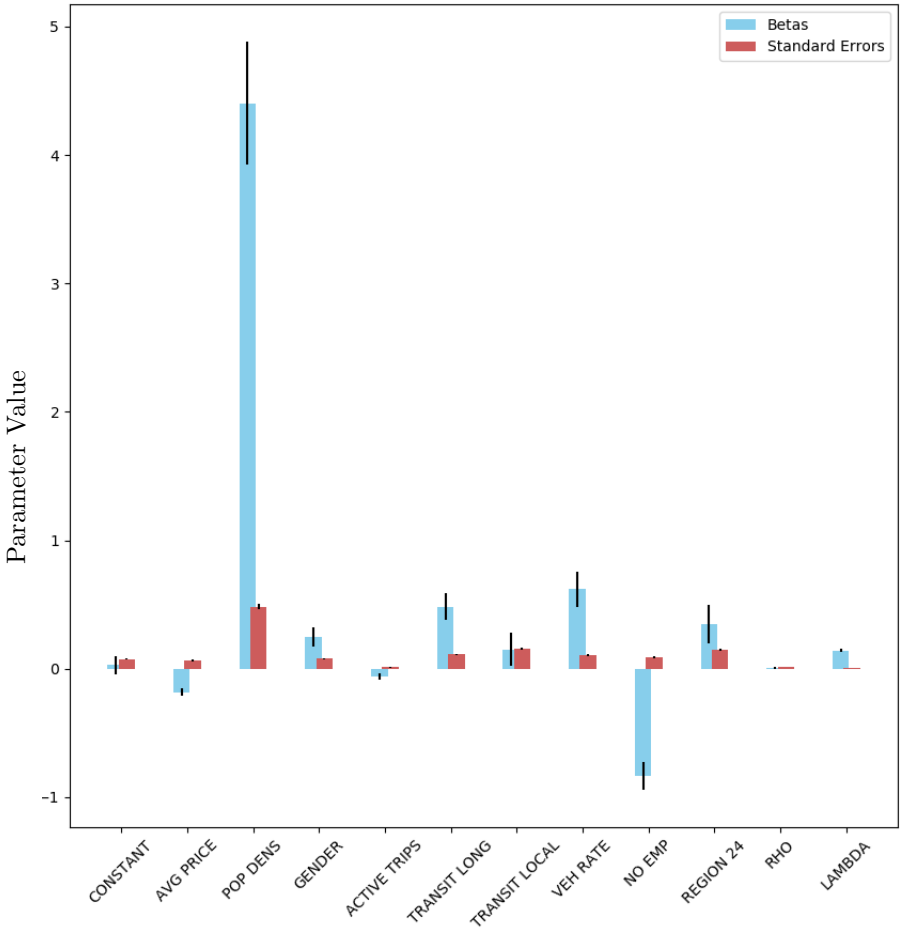

Figure 2: 100 bootstrap samples for total travel distance.

5.2 Work travel distance

Result for work travel distance are quite similar to those for total travel distance, suggesting a strong correlation between commuting distance and overall travel distance. In the case of commuting distance, the region-specific parameter for Durham comes out stronger than the other four regions. This suggests that the lack of local employment in this region tends to increase commuting distances for residents relative to all other regions. We explored the distribution of land use between types (e.g. residential, commercial, industrial, institutional) for all models and find that industrial use within the TAZ tends to decrease commuting travel distance. However, this may be simply a function of there being few homes in TAZ with industrial land uses and therefore few TTS records. The LM test statistics for work travel distance give a similar result to those of the total travel distance model.

Application of the model to commuting trips only suggests the previous hypothesis about the sign for population density by Morency et al. [23] is incorrect. While it is true that non-motorized trips are under reported in travel surveys, it is highly improbable that respondents are not reporting their own commuting trips, regardless of the travel mode. This suggests that other factors are contributing to this unexpected sign. It is likely an issue of an unobserved variable that simultaneously affects both travel distance and population density.

Regarding the bootstrapping exercise, the parameters are similarly invariant as in total travel distance (see Figure 3). The parameter magnitudes are also quite similar.

Table 3: Summary of results for work travel distance

Variable	Param	Std Err.	t-stat	p-value
Constant	-0.09	0.084	-1.117	0.264
Population density of TAZ (persons per km^2)	3.11	0.306	10.149	0.000
Gender	0.47	0.108	4.374	0.000
Active trip count	-0.13	0.018	-7.04	0.000
Transit pass (GO/PRESTO)	0.77	0.137	5.613	0.000
Transit pass (local)	0.06	0.205	0.278	0.781
Vehicles per driver	0.70	0.137	5.109	0.000
No employment	-0.61	0.215	-2.82	0.005
Home in Durham region	0.74	0.195	3.814	0.000
Home in Halton, Peel, or York region	0.32	0.151	2.101	0.036
Home TAZ percent industrial	-1.10	0.38	-2.893	0.004
ρ	0.00		0.009	
λ	0.20		0.007	
Log likelihood	-3739.9			
AIC	7505.8			
Number of observations	979			
Pseudo R-squared	0.164			

LM Test Statistics

H_0	χ^2	p-value
Joint $\rho = 0$ and $\lambda = 0$	292.1	0
Joint $\rho = 0$ and $\lambda = 1$	99807	0
1D $\rho = 0 \vert \lambda = 0$	290.4	0
1D $\rho = 0 \vert \lambda = 1$	99807	0
1D $\lambda = 0 \vert \rho = 0$	290.8	0
1D $\lambda = 1 \vert \rho = 0$	99807	0
Cond. $\rho = 0$ and λ unknown	1.5	0.215
Cond. ρ unknown and $\lambda = 0$	297.6	0
Cond. ρ unknown and $\lambda = 1$	100075.7	0

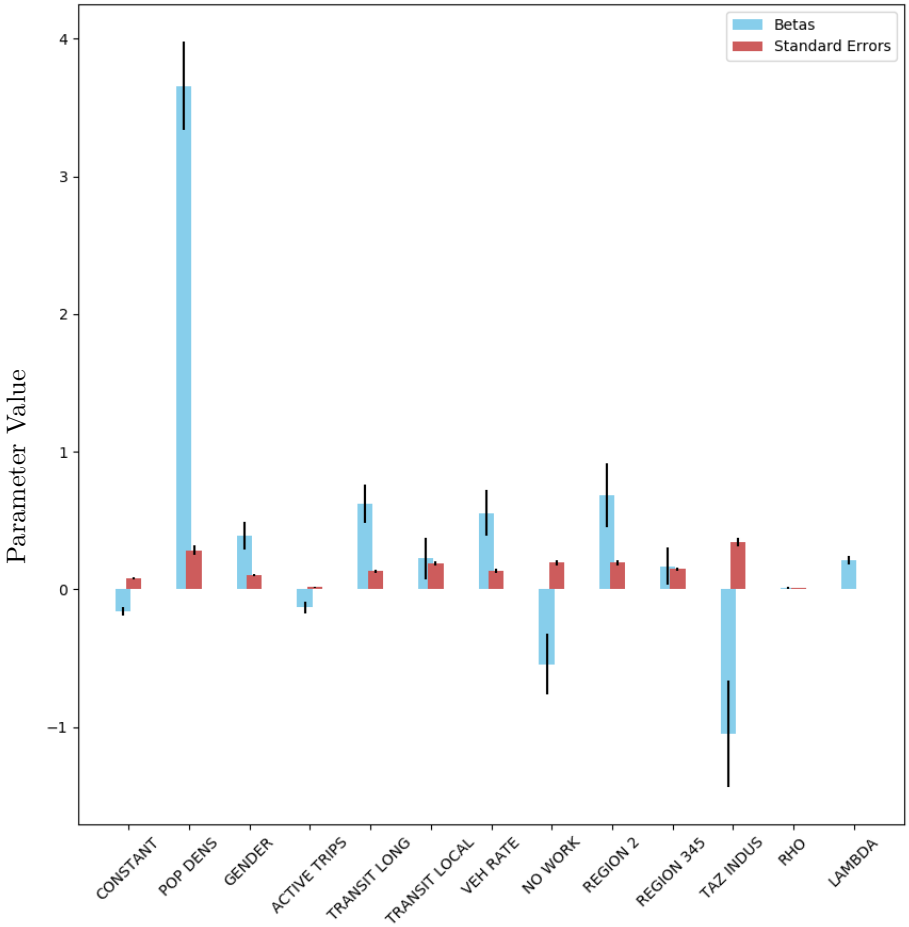

Figure 3: 100 bootstrap samples for work travel distance.

5.3 Shopping travel distance

The shopping travel distance model does not include as many significant variables as the previous two models. We include household size in this model because this will affect the amount of shopping required by the household.

Again, the bootstrapping exercise indicates quite robust parameters and that spatial dependence is minimal in all 100 samples.

Table 4: Summary of results for shopping travel distance

Variable	Param	Std Err.	t-stat	p-value
Constant	-0.03	0.08	-0.337	0.736
Population density of TAZ (persons per km^2)	1.1	0.291	3.792	0.000
HH size	0.17	0.045	3.724	0.000
Active trip count	-0.05	0.011	-4.156	0.000
Vehicles per driver	0.28	0.14	2.029	0.042
Home in Durham or York region	0.26	0.173	1.524	0.127
ρ		0		0.005
λ		0		0.024
Log likelihood		-1421.5		
AIC		2859.1		
Number of observations		460		
Pseudo R-squared		0.035		

LM Test Statistics		
H_0	χ^2	p-value
Joint $\rho = 0$ and $\lambda = 0$	2.4	0.306
Joint $\rho = 0$ and $\lambda = 1$	69924.8	0
1D $\rho = 0 \vert \lambda = 0$	1.8	0.176
1D $\rho = 0 \vert \lambda = 1$	69924.4	0
1D $\lambda = 0 \vert \rho = 0$	2.4	0.124
1D $\lambda = 1 \vert \rho = 0$	69924.7	0
Cond. $\rho = 0$ and λ unknown	0	0.855
Cond. ρ unknown and $\lambda = 0$	2.4	0.123
Cond. ρ unknown and $\lambda = 1$	69929.5	0

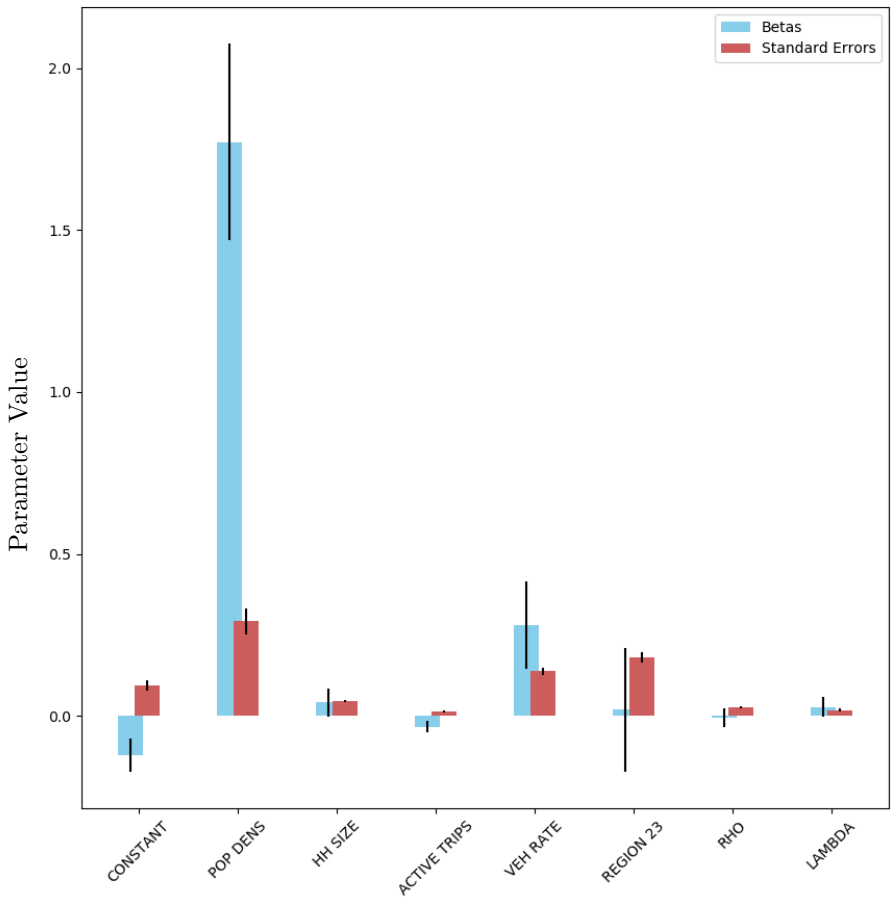

Figure 4: 100 bootstrap samples for shopping travel distance.

5.4 Validation of Spatial Lag Findings

The lack of spatial dependence in travel distance was an unexpected result. We believe the above analysis provides strong evidence for this outcome, but wanted to explore it further using another spatial statistics technique. Local indicators of spatial association (LISA) is a technique that works quite well for examining spatial dependence between regions. It is based on the calculation of a measure of deviation from a global average and identifying adjacent regions that exhibit significant variation from the mean. In the present instance, we calculate the average travel distance for each TAZ as the average of TTS respondents residing in each TAZ (using the full TTS dataset). A map can be developed using the LISA statistics, wherein the average travel distance for each TAZ is compared against the global mean for the GTA. There are 4 categories in the LISA map presented in Figure 5: HH (TAZ has an average travel distance above the global mean and its adjacent TAZ also have travel distances above the global mean), LH (TAZ has an average travel distance below the global mean and its adjacent TAZ have travel distances above the global mean), HL (TAZ has an average travel distance above the global mean and its adjacent TAZ have travel distances below the global mean), LL (TAZ has an average travel distance below the global mean and its adjacent TAZ have travel distances below the global mean)). For each case, the classification is compared against a 0.01 significance level. Tests can be performed using simulation to determine whether the pattern of classifications differs from that expected through random chance.

Figure 5 supports the hypothesis that TAZ located further from the Lake Ontario shoreline and outside the City of Toronto tend to have longer travel distance. TAZ located in the western portion of the City of Toronto tend to have shorter travel distances, supporting the importance of transit, density, and walkability on reduced travel distance (these are dense and walkable areas of the city). Regarding spatial dependence, this analysis supports the lack of spatial dependence. First, there are large regions that do not differ significantly

from the global mean travel distance. Second, most of the densely populated City of Toronto (i.e. high response density in TTS) have uniformly low travel distances and do not exhibit high variation between TAZ. Finally, the LISA map uses the average for each TAZ and variation also exists between respondents within each TAZ. Spatial lag dependence arises from a combination of high variation between TAZ, on the one hand, and low variation within each TAZ on the other. Any variation in the intra-TAZ travel distances will only exacerbate the effect of the low inter-TAZ variations.

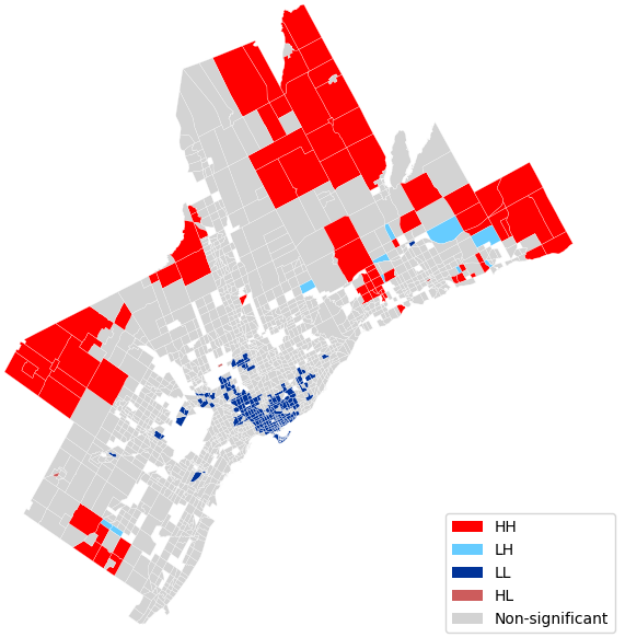

Figure 5: LISA map for average travel distance by TAZ.

6 Conclusions and Future Work

Finalize methods/results with Habib, then write this section.

References

[1] Paulus Teguh Aditjandra, Corinne Mulley, and John D Nelson. "The influence of neighbourhood design on travel behaviour: Empirical evidence from North East England". In: *Transport Policy* 26 (2013), pp. 54–65. DOI: 10.1016/j.tranpol.2012.05.011. URL: http://dx.doi.org/10.1016/j.tranpol.2012.05.011.

[2] Luc Anselin. *Spatial Econometrics: Methods and Models*. Springer Netherlands, 1988. DOI: 10.1007/978-94-015-7799-1.

[3] Luc Anselin, Serge Rey, and Levi Wolf. *Python Spatial Analysis Library*. 2018. URL: https://pysal.readthedocs.io/en/latest/library/spreg/ml_lag.html.

[4] Anzhelika Antipova, Fahui Wang, and Chester Wilmot. "Urban land uses, socio-demographic attributes and commuting: A multilevel modeling approach". In: (2011). DOI: 10.1016/j.apgeog.2011.02.001. URL: www.crpc-la.org/crpc/Documents/UPWP%2520FY%.

[5] Badi H Baltagi and Dong Li. *Test for Linear and Log-Linear Models against Box-Cox Alternatives with Spatial Lag Dependence*. 2004. DOI: 10.1016/S0731-9053(04)18001-8. URL: https://doi.org/10.1016/S0731-9053.

[6] Canh Quang Le and Dong Li. "Double-Length Regression Tests for Testing Functional Forms and Spatial Error Dependence". In: *Economics Letters* 101 (2008), pp. 253–257.

[7] Emilio Casetti. "Generating Models by the Expansion Method: Applications to Geographical Research". In: *Geographical Analysis* (1972). ISSN: 15384632. DOI: 10.1111/j.1538-4632.1972.tb00458.x.

[8] Pierre A Chiappori et al. "Couple Residential Location and Spouses Workplaces". In: *III Workshop on Urban Economics, June 9-10, Barcelona* (2014), p. 34. URL: http://www.sustaincity.org/publications/WP_3.3_Couple_Chaippori_de_Palma_Picard.pdf%20http://www.ieb.ub.edu/files/PapersWSUE2014/Inoa.pdf.

[9] Russell Davidson and James G Mackinnon. "Artificial Regressions". URL: https://pdfs.semanticscholar.org/6eea/52a0e46ff1fc3b7a29e4fc1f447633dc7db3.pdf.

[10] Russell Davidson and James G Mackinnon. *Double-Length Artificial Regressions*. Tech. rep. 1988, pp. 203–217. URL: http://qed.econ.queensu.ca/working_papers/papers/qed_wp_691.pdf.

[11] João De Abreu E Silva, Catherine Morency, and Konstadinos G Goulias. "Using structural equations modeling to unravel the influence of land use patterns on travel behavior of workers in Montreal". In: *Transportation Research Part A* 46 (2012), pp. 1252–1264. DOI: 10.1016/j.tra.2012.05.003. URL: http://dx.doi.org/10.1016/j.tra.2012.05.003.

[12] Chuan Ding et al. "Influences of built environment characteristics and individual factors on commuting distance: A multilevel mixture hazard modeling approach". In: (2017). DOI: 10.1016/j.trd.2017.02.002. URL: http://dx.doi.org/10.1016/j.trd.2017.02.002.

[13] Erik Elldér. "Residential location and daily travel distances: the influence of trip purpose". In: *JOURNAL OF TRANSPORT OF GEOGRRAPHY* 34 (2014), pp. 121–130. DOI: 10.1016/j.jtrangeo.2013.11.008. URL: http://dx.doi.org/10.1016/j.jtrangeo.2013.11.008.

[14] Erick Guerra et al. "Residential location, urban form, and household transportation spending in Greater Buenos Aires". In: (2018). DOI: 10.1016/j.jtrangeo.2018.08.018. URL: https://doi.org/10.1016/j.jtrangeo.2018.08.018.

[15] Kent M Hymel. *Factors Influencing Vehicle Miles Traveled in California: Measurement and Analysis*. Tech. rep. California State University, 2014. URL: https://sor.senate.ca.gov/sites/sor.senate.ca.gov/files/ctools/CCS_Report--Factors_Influencing_Vehicle_Miles_Traveled_in_California.pdf.

[16] Dena Kasraian, Kees Maat, and Bert Van Wee. "Urban developments and daily travel distances: Fixed, random and hybrid effects models using a Dutch pseudo-panel over three decades". In: *Journal of Transport Geography* 72 (2018), pp. 228–236. DOI: 10.1016/j.jtrangeo.2018.09.006. URL: https://doi.org/10.1016/j.jtrangeo.2018.09.006.

[17] James P Lesage. *The Theory and Practice of Spatial Econometrics*. 1999.

[18] Dong Li and Canh Le. "Nonlinearity and Spatial Lag Dependence: Tests Based on Double-Length Regressions". In: *Journal of Time Series Econometrics* 2.1 (2010). DOI: 10.2202/1941-1928.1039.

[19] Kevin Manaugh et al. "The effect of neighbourhood characteristics, accessibility, home-work location, and demographics on commuting distances". In: *Transportation* 37 (2010), pp. 627–646. DOI: 10.1007/s11116-010-9275-z. URL: https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/00494488/v37i0004/627_teoncaladocd.xml.

[20] M Manoj and Ashish Verma. "Effect of built environment measures on trip distance and mode choice decision of non-workers from a city of a developing country, India". In: (2016). DOI: 10.1016/j.trd.2016.04.013. URL: http://dx.doi.org/10.1016/j.trd.2016.04.013.

[21] Jeffery Memmott. *Trends in Personal Income and Passenger Vehicle Miles*. Tech. rep. Bureau of Transportation Statistics, 2007. URL: https://www.bts.gov/sites/bts.dot.gov/files/legacy/publications/special_reports_and_issue_briefs/special_report/2007_10_03/pdf/entire.pdf.

[22] Luis F Miranda-Moreno et al. "Simultaneous Modeling of Endogenous Influence of Urban Form and Public Transit Accessibility on Distance Traveled". In: (). DOI: 10.3141/2255-11. URL: https://journals-sagepub-com.myaccess.library.utoronto.ca/doi/pdf/10.3141/2255-11.

[23] Catherine Morency et al. "Distance traveled in three Canadian cities: Spatial analysis from the perspective of vulnerable population segments". In: *Journal of Transport Geography* 19 (2011), pp. 39–50. DOI: 10.1016/j.jtrangeo.2009.09.013. URL: http://www.cimtu.qc.ca/EnqOD/Index.asp.

Appendices

A Hessian Matrix for General Box-Cox Model with Spatial Dependence

$$\frac{\partial^2 LL}{\partial\beta\partial\beta'} = \frac{-1}{\sigma^2}X'X \tag{A.1}$$

$$\frac{\partial^2 LL}{\partial\beta\partial\rho'} = \frac{-1}{\sigma^2}X'Wy^{(\lambda)} \tag{A.2}$$

$$\frac{\partial^2 LL}{\partial\beta\partial\lambda'} = \frac{1}{\sigma^2}X'(I-\rho W)C(y,\lambda) \tag{A.3}$$

$$\frac{\partial^2 LL}{\partial\beta\partial\sigma^{2\prime}} = 0 \tag{A.4}$$

$$\frac{\partial^2 LL}{\partial\rho\partial\rho'} = \frac{-1}{\sigma^2}(Wy^{(\lambda)})'(Wy^{(\lambda)}) - T_1 - T_2 \tag{A.5}$$

$$\frac{\partial^2 LL}{\partial\rho\partial\lambda'} = \frac{1}{\sigma^2}y^{(\lambda)}W'(I-\rho W)C(y,\lambda) + \frac{1}{\sigma^2}\hat{e}'WC(y,\lambda) \tag{A.6}$$

$$\frac{\partial^2 LL}{\partial\rho\partial\sigma^2} = \frac{-T_3}{\sigma^2} \tag{A.7}$$

$$\frac{\partial^2 LL}{\partial\lambda\partial\lambda'} = \frac{-1}{\sigma^2}[(I-\rho W)C(y,\lambda)]'[(I-\rho W)C(y,\lambda)] - \frac{1}{\sigma^2}\hat{e}'(I-\rho W)\partial C(y,\lambda) \tag{A.8}$$

$$\frac{\partial^2 LL}{\partial\lambda\partial\sigma^{2\prime}} = \frac{1}{(\sigma^2)^2}\hat{e}'(I-\rho W)C(y,\lambda) \tag{A.9}$$

$$\frac{\partial^2 LL}{\partial\sigma^2\partial\sigma^{2\prime}} = \frac{n}{2(\sigma^2)^2} \tag{A.10}$$

where (A.11)

$$\hat{e} = (I-\rho W)y^{(\lambda)} - X\beta \tag{A.12}$$

$$C(y,\lambda) = \frac{\partial C(y,\lambda)}{\partial\lambda} = \frac{\lambda y^{\lambda}(\ln y) - y^{\lambda} + 1}{\lambda^2} \tag{A.13}$$

$$\partial C(y,\lambda) = \frac{\partial C(y,\lambda)}{\partial\lambda} = \frac{\lambda^2 y^{\lambda}(\ln y)^2 - 2\lambda y^{\lambda}\ln y + 2y^{\lambda} - 2}{\lambda^3} \tag{A.14}$$

$$T_1 = tr[(W(I-\rho W)^{-1})^2] \tag{A.15}$$

$$T_2 = tr[(W(I-\rho W)^{-1})'(W(I-\rho W)^{-1})] \tag{A.16}$$

$$T_3 = tr[W(I-\rho W)^{-1}] \tag{A.17}$$