# An Analysis of Supervised Machine Learning Models for Road Traffic Forecasting.

William James Ashford

This dissertation is submitted in part requirement for the MSc in the Centre for Advanced Spatial Analysis, Bartlett Faculty of the Build Environment, UCL.

Supervisor: Mr Steven James Gray

BENVGSC6

August 2017

Word Count: 11858

# Abstract

The accurate forecasting of road traffic is an important pursuit in statistical and computational sciences as it shows promise to provide citizens, governments, and other interested stakeholders with information regarding future traffic congestion events. With this information, it is envisaged that the negative externalities associated with traffic congestion may be better minimized, namely economic losses due to delays and air pollution.

In recent decades, there has been a growth in the application of spatiotemporal machine learning models to traffic forecasting in the literature (Vlahogianni, et al., 2014). The status quo of the literature in representing developments in machine learning traffic forecasting models has been to iteratively adjust characteristics of models, known as hyperparameters, through a process of trial and error. The results of a handful of these models are then typically compared in a process akin to performing t-tests with a sample size of near to one (Castro-Neto, et al., 2009) (Lv, et al., 2015) (Kingma & Ba, 2015) (Liu, et al., 2011) (Min & Wynter, 2011) (Moretti, et al., 2015) (Yasdi, 1999).

If research for road traffic forecasting models is to be of benefit to practitioners, the results must be reliable and as it stands, the literature is not providing practitioners with reliable findings.

The aim of this paper is to contribute to the understanding of the nature of optimal road traffic forecasting models for the general case to better inform practitioners. It sets a precedent for the presentation of reliable results in the field through

development of a randomised hyperparameter allocation algorithm, through which the performance of different multi-layer perceptron artificial neural networks (ANNs) and support vector machines (SVMs) were compared. This algorithm was run in parallel on a 32-core machine, from which 186880 traffic forecasting models were developed and tested. K-means clustering of these models with respect to their hyperparameters and reporting the median performance of the cluster facilitated an analysis of the role of quantitative and qualitative hyperparameters in the prediction of traffic flow along all roads within the study sample, which consisted of 20 randomly selected roads in Aarhus, Denmark with data from 13 weeks in 2014.

ANNs were found to have lower error, on average, that SVMs. For ANNs, the L-BFGS learning algorithm showed the best performance with all activation functions performing as well as one another for this model species, except for the hyperbolic tangent function which outperformed the rectified linear function. For SVMs, the linear kernel outperformed all others.

Through the development of a binned minimum plot, trends in hyperparameters were able to be understood in the case that other hyperparameters were optimised. This led to insights into the relationships between model complexity and error, the need for sufficient learning iterations, and into the relationships between model error and both momentum and learning rates in the case of ANNs and model error and regression boundary widths ($\varepsilon$) in the case of SVMs.

If the data used is representative of traffic generally, the findings of this study pertain to the general case.

## Declaration

I hereby declare that this dissertation is all my own original work and that all sources have been acknowledged. It is 11858 words in length.

Signed:

William James Ashford, August 2017.

Table of Contents

# Figures and Tables

# Introduction

The accurate forecasting of road traffic is an important pursuit in statistical and computational sciences as it shows promise to provide citizens, governments, and other interested stakeholders with information regarding future traffic congestion events. With this information, it is envisaged that the negative externalities associated with traffic congestion may be better minimized.

The aim of this paper is to contribute to the understanding of the nature of optimal road traffic forecasting models for the general case to better inform practitioners. In this context, the general case refers to traffic along any road at any time. Road forecasting models are functions typically comprised of several characteristics, known as hyperparameters, which define how they forecast traffic based on historical data.

Traditionally, as seen in the literature review, practitioners and researchers have developed traffic forecasting models through iteration of hyperparameters, which have been selected based on domain expertise or intuition.

It is the approach of this paper to conduct a robust statistical analysis of the relationships between model error and hyperparameters for several model types. From the results of this analysis, practitioners may derive an informed opinion of what quantitative and qualitative hyperparameters are likely to have low error and which initial values might be suitable for the commencement of a model tuning process.

The scope of this this paper is to analyse the hyperparameters of two major model classes, henceforth referred to as model genera. These model genera are artificial neural networks (ANNs) and support vector machines (SVMs) and are subsequently broken down into model species and subspecies. This nomenclature has been implemented to aid the reader in following the experimental methodology, results, analysis and discussion that will be covered in later sections.

Prior to explaining the grounds on which these model genera are further classified into model species and sub-species, and a technical introduction to their nature and the field more generally is required. Following this technical introduction, an introduction to the overall process of the analysis and nuances of the dataset will be introduced.

## Technical Introduction

The two main categories into which time series forecasting models are commonly grouped are statistical and machine learning models. While machine learning lies within the domain of statistics, the former specifically refers to models such as auto-regressive moving average models, simple identity models, Fourier transforms and spatiotemporal variations thereof. Machine learning models, which have received comparatively less attention in the literature (Vlahogianni, et al., 2014), include ANNs and SVMs, naïve Bayesians, random forests approaches, clustering algorithms and kernel ridge regression machines.

Both model genus reviewed in this study are known as supervised machine learning models in the literature.

Supervised machine learning is a process of developing predictive functions or 'models' through the fitting of that function to a given set of data, such that:

$$f(\vec{x}) = y$$

Whereby $\vec{x}$ is some input vector and $y$ is some output. The process of learning for these models is a process of adjusting the function to fit samples of training data, with the aim to arrive at a function that can produce the correct output for a new and unknown set of inputs (LeCun, et al., 1998). These outputs can be qualitative, in which case the model is known as a classification model, or they can be quantitative, in which case they are known are regression models. Regressive variants of ANNs and SVMs are the focus of this study.

SVMs and ANNs are fundamentally different model types, and as such, it is important to understand their similarities and differences.

*Artificial Neural Networks*

While there are a variety of ANN types, the focus of ANNs in this paper will be a variety known as multi-layer perceptron ANNs. These ANNs are a class of supervised machine learning model inspired by the function of biological brains. Following a fitting process, these models process an input vector via weighted links or synapses to several layers, each of which consists of a given number of neurons and provide an output. The typical neuron will process the sum of the products of input values through some activation function, such that:

$$f(\vec{x}, \vec{w}, b) = y$$

Where $\vec{x}$ is the input vector, $\vec{w}$ is the weights vector, representing the weights of the synapses upon which information has *travelled to* the neuron from the previous layer and $y$ is the output of that neuron and $b$ is a constant bias.

These neurons, layers and synapses can be combined in an infinite number of combinations to produce a neural network. The fitting process that enables a base architecture of layers, neurons and randomly weighted synapses to produce outputs of a given degree of accuracy is known as the learning function.

*Learning Function*

The learning function is an iterative process of adjusting synapse weights to minimise the error of the model. Weights are commonly adjusted through a process known as gradient-based learning (LeCun, et al., 1998), which attempts to find the gradient closest to zero within the hyperparameter-cost space. This requires a process to compute the partial derivatives of the total model error with respect to a parameter (i.e. a synapse weight) at a given neuron within the network and adjust parameters at that neuron such that, over subsequent iterations, the error of outputs for the entire network is reduced. To prevent overfitting, a parameter, $\alpha$, is used such that the product of the synaptic weight and $\alpha$ is added to any changes to that synaptic weight (LeCun, et al., 1998).

The process of identifying the partial derivatives of error with respect to specific synapses begins in each iteration by calculating the partial derivatives of total error with respect to synaptic weights feeding into the output node (in the case of a regressive ANN with a scalar output). Using the chain rule and the summation of lower order partial derivatives, the partial derivatives of the synapses that are directed towards these final synapses (for which partial error derivatives were computed) can also be computed (LeCun, et al., 1998).

In other words, the contribution of each synapse to the total error of the ANN is computed by first identifying the contribution of weights closest to the output. Following this, by understanding the contribution that an *upstream* synapse, $s_u$, has on the error of a *downstream* synapse, $s_d$, for which the contribution to the total model error has been computed, the contribution of $s_u$ to the total model error can be computed. This is because the sum of contributions to total error due to all $s_u$ synapses is equal to the contribution due to of $s_d$.

This process of identifying the partial derivatives, or error gradients, for synapse weights throughout an ANN is known as the back propagation of errors, or simply backpropagation (LeCun, et al., 1998) (Buscema, 1998).

Back-propagation, is not itself the learning function, but rather, is a process that provides a learning function with the error gradient and second order gradients of each synaptic weight which the learning function subsequently attempts to optimise.

Stochastic gradient descent (SGD) is a learning process whereby synapse weights are adjusted on the basis of the partial derivatives of error for each synapse. Finding the partial derivative of the error with respect to the weights (via backpropagation) and multiplying it by a learning rate, η, the new weights are computed, such that:

$$W_i(t) = W_i(t+1) - \eta \frac{\partial E}{\partial W_i}$$

SGD is commonly considered to be the default unconstrained optimisation algorithm that is well suited to ANN learning and against which, alternative unconstrained optimisation algorithms are compared (LeCun, et al., 1998) (Kingma & Ba, 2015). Several software tools also facilitate error dependent learning rates, such as rules that the rate reduces depending on stagnation in performance (Pedregosa, et al., 2011) (Baydin, et al., 2017).

ANNs are susceptible to local minima problems, and as such, the concept of momentum can be applied to SGD ANNs (Rumelhart, et al., 1986). Momentum allows for some proportion of the previous change in synaptic weight to carry over to the current iteration, such that,

$$W_i(t) = W_i(t+1) - \eta \frac{\partial E}{\partial W_i} - m\Delta W(t-1)$$

A further development of the SGD learning function is Adam [sic] (Karlaftis & Vlahogianni, 2011). Adam is a stochastic optimisation algorithm similar to SGD, however instead of a fixed learning rate or a learning rate that adapts in the event of

stagnation, a continually adaptive learning rate is used (Kingma & Ba, 2015). This learning rate is adapted based on two 'moments', $m_t$ and $v_t$, of the gradient of the cost function, equal to the partial derivative used in SGD, such that:

$$m_t = \beta_1 m_{t+1} + (1 - \beta_1) * \frac{\partial E}{\partial W_i}$$

and,

$$v_t = \beta_2 v_{t+1} + (1 - \beta_2) * \left(\frac{\partial E}{\partial W_i}\right)^2$$

Where $\beta_1$and $\beta_2$ are rates that govern the decay of these moments. Given the algorithm decays $m_t$ and $v_t$ in such a manner that near zero values are biased, the moments are balanced (Kingma & Ba, 2015), such that:

$$\widehat{m_t} = \frac{m_t}{1 - \beta_1}$$

and,

$$\widehat{v_t} = \frac{v_t}{1 - \beta_2}$$

These balanced moments are combined to amend the SGD's functional formula when applied to ANN synaptic weights such that:

$$W_i(t) = W_i(t+1) - \eta \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon}$$

Where $\epsilon$ is a constant (Kingma & Ba, 2015).

In each of these learning algorithms, the learning rate is instrumental in adjusting synapse weights. In the case of generalised traffic forecasting models, high learning rates for more accurate models might indicate that the spatiotemporal relationships for traffic flow evolve and that traffic events closer together in time share similar underlying dynamics. Lower learning rates, however, might indicate that the underlying dynamics for spatiotemporal relationships in traffic flow are more consistent over time, especially for a low number of maximum training iterations.

SGD and Adam are optimisation algorithms based on the rate of change in error with respect to synaptic weights. The Limited Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) is a second order optimisation, one that is based on the rate of change of the rate of change in error with respect to synaptic weights.

L-BFGS technique which is similar to the iterative Newton's method. In Newton's method, a matrix of all second order partial derivatives, known as the Hessian, is combined with the cost gradient of the function to iterate towards an optimal combination of independent variables for a given function. In the case of optimising a neural network, these variables are the synaptic weights.

Newton's method requires computation of $n^2$ second partial derivatives to construct the Hessian, where n is the number of synaptic weights. Rather than compute the $n^2$ second partial derivatives, the Hessian can be approximated through a function, known as a Quasi-Newton.

The BFGS is one such approximation, which is a function of the historical changes in the gradient of the cost function and the parameters (i.e. synaptic weights). To preserve memory in the case of an ANN that requires an exorbitant amount of iterations to reach a minimum of cost, the number of historical samples is limited to *m* past iterations; hence, Limited-BFGS (Pedregosa, et al., 2011).

These three activation functions represent three approaches to the unbound optimisation problem that is ANN learning. SGD is a popular first order approach Adam is modern and popular adaptation of SGD and L-BFGS is an appropriate second order approach.

Given these algorithms govern how an ANN learns, in this study they are designated as the characteristics that distinguish model species from one another within the ANN model genus.

*Activation Functions*

The activation function of a neuron in an ANN is the function that processes the layer inputs and synapse weights into a neuron into a given output. Some commonly applied activation functions include:

- The identity function, $f(\vec{x}, \vec{w}) = \vec{x} \cdot \vec{w}$,
- The logistic function, $f(\vec{x}, \vec{w}) = \frac{1}{1+e^{-\vec{x} \cdot \vec{w}}}$,
- The hyperbolic tangent function or *tanh*, $f(\vec{x}, \vec{w}) = tanh(\vec{x} \cdot \vec{w})$,
- The rectified linear unit function or *relu*, $f(\vec{x}, \vec{w}) = max(0, \vec{x} \cdot \vec{w})$.

The number of possible activation functions is infinite. As such an experiment reviewing just the aforementioned functions would be non-exhaustive and has the possibility to produce local maxima for model accuracy within the universal hyperparameter space.

Neuron activation functions is the qualitative hyperparameter that is used to identify ANN model sub-species examined in this study.

*Support Vector Machines*

SVMs are models that were initially developed to classify input vectors into one of two categories (positive or negative) through the optimisation of a linear margin the separates the data points belonging to each category (Boser, et al., 1992).

The margin is generated through a Lagrangian optimisation process which is dependent on all the dot products of all the possible pairs of samples. Standard linear SVMs may fail to converge when the data is not linearly separable, and as such, transformation of samples into a higher dimensionality feature space is undertaken and the dot product of all pairs of samples are found within this new feature space (Boser, et al., 1992). Conceptually, it is in this feature space that the data become linearly separable and the subsequent optimisation of the margin occurs.

Transformation to high dimensionality feature spaces for large datasets may prove too computationally expensive for some scenarios, and as such, the concept of a kernel is often applied. A kernel is a function that returns the dot product of two vectors (i.e.

samples) from a feature space without the need to transform the sample vectors to that feature space and are computationally inexpensive relative to the alternative.

For the case of linearly inseparable samples, an SVM conducts a transformation n-dimensional input vector, $\vec{x}$, using a function known as a kernel, which returns the dot product of the vectors in the alternative dimensionality space.

SVMs can be used for regression by adapting the classification process such that the locus of a margin of predefined width 2ε is such that the margin contains all samples within the higher dimensionality feature space. A misclassification penalty, C, determines the extent to which deviations greater than epsilon are permitted (Smola & Schölkopf, 2004).

The kernel functions examined within this study are:

- Linear: $K(\bar{x}_i, \bar{x}_j) = \bar{x}_i{}^T \bar{x}_j$
- Radial Base Function: $K(\bar{x}_i, \bar{x}_j) = e^{\gamma(||\bar{x}_i - \bar{x}_j||^2)}$

  Where $\gamma$ is a constant.
- Polynomial: $K(\bar{x}_i, \bar{x}_j) = (\bar{x}_i \cdot \bar{x}_j + c)^d$

  Where c is a constant and d is the degree of the polynomial kernel.
- Sigmoid: $K(\bar{x}_i, \bar{x}_j) = \tanh(\gamma(\bar{x}_i \cdot \bar{x}_j) + c)$

  Where c and $\gamma$ are constants.

For the model genus of SVMs, species were distinguished from one another based the kernel function that was applied.

## Dataset Overview

The dataset examined in this study consist of 111.4 days of traffic count data from 2014 (32076 five-minute period) from the urban road network of Aarhus, Denmark and reports on 449 roads within and around the city. The dataset contains the distance of the 449 roads, speed limit of the roads, as well as traffic flow, mean speed, median speed and mean measured travel times for each five-minute period.

The data was collected via a network of Bluetooth sensors and registered the passage of cars with Bluetooth signatures (Bloksgaard & Christiansen, 2015).

Of note is that the passage of non-Bluetooth compatible vehicles is not recorded with this system. An assumption of this study is that the ratio of Bluetooth compatible to non-Bluetooth compatible vehicles is spatiotemporally homogenous. As such, the data is treated as consistently representative of the total traffic flow.

The network is a non-exhaustive representation of the physical road network it represents, as seen in figure 1.


*Figure 1 The extent of the of the Bluetooth sensor network, as presented in the original paper (Bloksgaard & Christiansen, 2015). Note the roads in the centre of Aarhus without sensors.*

## Overall Process

The experimental process of this study was designed with time and resource constraints in mind. To generate sufficient data on the relationship between multiple hyperparameters and error, significant computation was required.

Utilising parallel computing in cloud environments facilitated a far greater scope of experimentation that would have been possible with conventional local machines, however, to prevent unviable costs, some adjustments were made to the extent of the experimentation. The extent of the experimentation was subsequently set at 20 pseudo-randomly selected roads from the dataset. The roads had to be predetermined and the randomisation algorithm run on those roads' data due to the memory requirements associated with loading training data for models. Justifications around this scope are detailed further in the methodology section.

The experimentation in this study can be represented in the flowchart in figure 2.

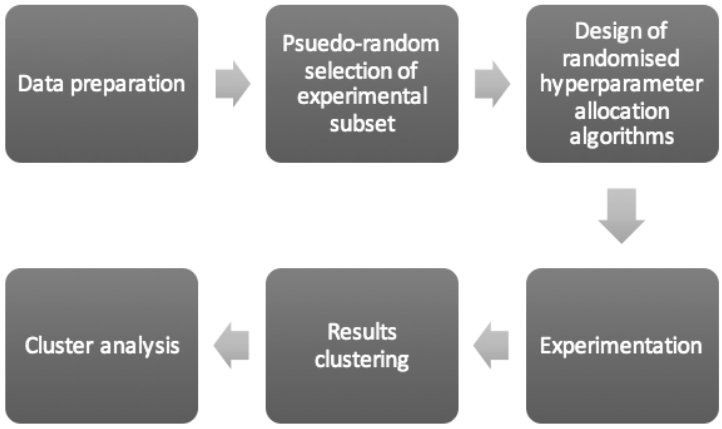



*Figure 2 A flow chart outlining the overall process of the study.*

# Literature Review

The development of forecasting models for road traffic in the literature is well established with a variety of model classes developed to forecast road traffic metrics (Vlahogianni, et al., 2014). These metrics include traffic volume, speed and travel duration (Vlahogianni, et al., 2014).

Models in the literature forecast a traffic metric at time *t* based on past records of that metric over a given number of temporal lags (i.e. at t – n). Alternative models also include historical data of neighbouring road segments, with reports of increased accuracy (Haworth, 2014) (Chandra & Al-Deek, 2009). Such models are known as spatiotemporal models.

This literature review will focus on the practices common in the development of traffic forecasting models and alternatives standard set out in the literature. Given SVM and ANN model genera are the focus of this study, the scope of this review is the development of these models for traffic forecasting applications.

What is seen throughout this literature is two common practices. The first is the ubiquitous application of hyperparameter tuning; a process of selecting initial hyperparameters and, through a process of trial and error, arriving at an acceptable model configuration. The second is the comparison of model genera, species or sub species based on the accuracy of a handful of models from each classification with an equally small diversity in hyperparameter values.

## Artificial Neural Networks

The application of ANNs to traffic flow forecasting has been investigated in the literature with several simple and complex models. Due to what is anticipated to be a lack of data and a lack of computational resources, the development of ANNs for traffic forecasting has focused on datasets with a small number of road sites with the tuning of model hyperparameters to arrive at a viable model for those road sites.

This was present in Yasdi's 1999 study which applied a partially recursive ANN to forecasting of traffic flows at one site along a German highway. The temporal granularity of the raw data in this study was 5 minutes but due to concerns of 'randomness' in the data and computational constraints, data was aggregated to '10 or 15 minute intervals' (Yasdi, 1999). In the process of configuring the ANN, the networks' architectures were 'carefully selected' to include temporal data, facilitate greater forecast horizons (i.e. the functionality to predict both $x_{t+1}$ and $x_{t+n}$) and, to minimise computational complexity. Yasdi's ANN was configured via trial and error as, at the time of writing, there were 'no systematic procedures to find the optimal configuration of a neural network' (Yasdi, 1999).

In the comparison of the accuracy of an ANN against a statistical model in the forecasting of traffic flows along three segments of freeway in Melbourne, Australia, a single model architecture and learning algorithm was used (Li & Rose, 2011). In the paper describing this model, there is no indication that consideration or testing of alternative ANN architectures, learning functions or other hyperparameters was undertaken.

In the development of an ensemble model for forecasting of traffic flow rates for three streets North of Rome, Moretti, et. al used a single layer ANN with eight input nodes (temporal lags), ten hidden nodes, sigmoid activation functions and a maximum iteration limit of $10^6$. (Moretti, et al., 2015). The only hyperparameter that was not selected via intuition was the number of temporal lags, which was selected via a tuning process.

Zhang, et. al., in their development of a Bayesian ensemble model, used a SGD ANN with set momentum, a single 20-neuron hidden layer, an initial learning rate of 1000 and a hyperbolic sigmoid activation function (Zheng, et al., 2006). While the intent of this paper was to perhaps demonstrate the synergistic effect of a Bayesian ensemble, the sample sizes were not sufficient and the hyperparameters of the ANN created in this way were, nonetheless, selected using the practitioners' intuition.

This trend of hyperparameter selection continues throughout the reviews of the state of the art of the application of ANNs traffic forecasting (Vlahogianni, et al., 2014) (Habtie, et al., 2017). While this practice may be expected commercially, this research does little to present statistically robust findings to practitioners.

## Support Vector Machines

The literature detailing research into the application of SVMs for traffic forecasting deviates little from the case of the ANN literature.

In the development of a regressive SVM for forecasting of highway traffic flow in Taiwan, Wu, et. al. followed a process of trial and error to arrive at a linear SVM with

a margin of 0.02 and C value of 1000 (Wu, et al., 2004). In the development of a SVM for the purposes of comparison to other model types, Castro-Neto, et. al. selected a RBF kernel, with no mention of consideration for other kernels (Castro-Neto, et al., 2009).

In the comparison of different model genera on the forecasting of traffic flow along Californian Highways, Lv, et. al. implement a linear SVM. No commentary was made on the nature of hyperparameters or their selection process (Lv, et al., 2015), as such, a standard linear SVM is assumed, however the epsilon and C values chosen are unknown, as is the method of their selection.

This is also the case in Ding, et. al.'s study of the application of regressive SVMs to traffic flow forecasting in the city of Xi'an, China (Ding, et al., 2002). The omission of details surrounding hyperparameter values in these studies demonstrates either a case of human error or the lack of value placed on hyperparameters of SVMs by the authors.

## Alternative Methods for Model Development

Perhaps the most advanced approach in the literature for the selection of a given models hyperparameters is that of a 'genetic' approach, as detailed in Yao & Lin's 1998 paper. This approach conducts a grid search on different hyperparameters and only updating the model's hyperparameters if a more accurate model is discovered in the grid search (Yao & Lin, 1998).

The process of genetic optimisation of machine learning models for the purposes of traffic forecasting has precedent in the literature (Vlahogianni, et al., 2014) (Vlahogianni, et al., 2004) (Adeli, 2001).

This process of model selection can be justifiable termed a black box; an approach that may well produce an optimal solution, but the explanatory power of the model is near non-existent. As such, these methods do little to aid heavily resource constrained practitioners who could benefit from reliable findings in the literature.

## Summary

To date, the development of machine learning models for traffic forecasting has shown promise and there have been a wide range of models that have been explored. Models range in their complexity, accuracy and spatiotemporal scope (Vlahogianni, et al., 2014) (Vlahogianni, et al., 2004).

The status quo for model development in the literature is predominantly one in which practitioners follow a trial and error process to arrive at an appropriate model. If the range of hyperparameters selected is poor enough, the practitioner risks producing a locally optimal model with performance potentially much lower than the globally optimal case. Moreover, the results are not a reliable indication of model genera, species and sub-species for the general case given their limited scope and small sample size.

When researchers have attempted to compare the viability of given genera, species or subspecies of model to one another, they have done so with a limited sample of the

populations on a frequently small number of roads or training samples (Castro-Neto, et al., 2009) (Lv, et al., 2015) (Kingma & Ba, 2015).

This is akin to conducting an unpaired t-test with a handful of individual samples. This is not a statistically robust way to assess the relative performance of model genera, species or subspecies in the general case and does little to aid practitioners.

Review of the literature reveals that the current state of research into globally optimal model genera, species and subspecies is at worst, a process of comparing a few creatively designed models for a small number of road segments and at best, an investigation of the output of genetic model development algorithms for a small number of roads (Vlahogianni, et al., 2014), (Vlahogianni, et al., 2004), (Adeli, 2001).

There have been no examples of reliable investigations into the nature of the relationships between hyperparameters or input variables and model accuracy in a statistical robust way. This is to the detriment of resource constrained practitioners who may not have the time, resources or skill to employ more robust hyperparameter sweeping or genetic algorithms, and who must select hyperparameters from spectra and incrementally adjust them through a process of trial and error.

This lack of reliable analysis is perhaps a result of the computational resources needed to complete such an analysis but also of the focus of the field generally. As noted in Adeli, 2001, the focus of ANNs in traffic engineering has been the forecasting of traffic metrics along a given road (Adeli, 2001). If research follows the same focus as

practice, models developed in research may be, from the outset, prone to a local overfitting as it is only the local instance that is of interest.

Research may yield greater benefit to practitioners if that research is conducted with a perspective focused on the general case. Adoption of such a perspective represents an untapped opportunity to not only develop more accurate traffic forecasting models but to potentially understand the nature of traffic for the general case.

# Methodology

This section outlines the methodology of this study, beginning with the selection of the forecasting metric, definitions of model accuracy, data preparation and experimental design.

## Forecasting Metrics

While contemporary models are frequently applied to the forecasting of either travel time or traffic flow (Vlahogianni, et al., 2014), this study was focused on the forecasting of traffic flow. This decision was based on nature of the collection and the subsequent recording of traffic metrics.

In the dataset, periods with a net flow of zero vehicles were allocated the travel duration of the most recent non-zero window. The flaw of this process was demonstrated upon review of several data whereby travel speeds exceeding 100km/h and the associated travel times were recorded alongside zero traffic flow following a rapid vehicle passage in the previous non-zero flow recording. This occurred on roads with much higher frequencies of low speed travel and was evidently an error in the recording process.

## Model Accuracy

In measuring model accuracy for these forecasting models, several metrics were considered; namely the mean absolute error (MAE), the root mean square error (RMSE), the median absolute error (MEDAE), the mean absolute percentage error (MAPE) and the correlation of determination ($R^2$).

Given the data contained zero values for periods where no vehicle travelled along a given road segment, MAPE was not a viable measure as it is the mean of the quotients of the absolute difference between the actual and predicted results and the actual result. An actual result of zero would result in a division by zero, thereby rendering an undefined APE for such a point.

$R^2$ was not an appropriate measurement as identical stepwise variance with different flow magnitude would return a perfect $R^2$ (Armstrong, 2001). RMSE is a metric heavily influenced by scale (Armstrong, 2001) and as such, was also not appropriate.

MAE is a popular error metric used in the literature and is not unacceptably influenced by scale. It is however, influence by scale to a degree and as such MEDAE was also used. For the purposes of the study, MEDAE was used as the primary error metric, whereby MAE was used to remain consistent with the literature and for reference.

## Data Preparation

In the pre-processing of the available data, road segments that contained exclusively null values or all zero values were removed from consideration. Whether these data resulted from collection errors or if they were simply roads along which no traffic travelled was irrelevant as the purpose of this study was to understand the relationship between hyperparameters and error. A complete zero set would not provide a valid canvas upon which these relationships could be tested.

In preparing the non-zero data for analysis, an additional pre-processing step undertaken was dimension-wise scaling of input vectors to values between 0 and 1. While frequently undertaken to reduce the impact of scale on outputs, given the fact that all dimensions in the input vectors pertained to traffic flow, the removal of scale dependence was not a primary concern. Rather, this standardisation was conducted to improve computational efficiency of the hyperparameter randomisation algorithm.

## Experimental Design

The experiment in this study was designed to facilitate a statistical analysis of relationships between hyperparameters and model accuracy independent of the road a model was developed on. As such, performance of models within a spectrum of hyperparameter values was required and a record of the accuracy of models with those hyperparameter values across a variety of roads was required.

To facilitate this, a randomised hyperparameter allocation algorithm was developed, in an approach akin to genetic development of ANNs by randomised evolution (Bergstra & Bengio, 2012), except that instead of providing the optimal model within the range of hyperparameters, it provided a summary statistics of models with randomised hyperparameters.

For each road examined, hyperparameters were randomly allocated to a model, the model was trained on the first 75% of the data (ordered chronologically) and was tested on the remaining 25%.

In allocating an appropriate number of randomised models to each genus, cost constraints, the computational limits of the virtual machines used and the prevalence of the local minima problem for ANNs (Boser, et al., 1992) were all considered. The resulting programmed ratio was 8.41 ANNs to 1 SVM, which resulted in the training and testing of 167040 ANNs and 19840 SVMs.

The development of this many models, far exceeding the numbers of those developed in the literature, was facilitated using parallel computation on a 32-core virtual machine available through the Amazon Web Services (AWS) cloud computing platform. Storage of results was facilitated using a common relational database.

These machines were preconfigured with the required software (Python) and associated packages (Pedregosa, et al., 2011) to facilitate the extraction, transformation and, manipulation of input data, the development of the models and the submission of results the database. The results recorded in the database detailed the MAE and MEDAE of models run on single roads, i.e. they were not representative of the general case.

To gain an understanding of the performance of similar models on different roads, a k-means clustering algorithm was employed with a cluster size of 10% of the number of rows pertaining to the specific model subspecies. The median MAE and MEDAE of each cluster was then recorded as the cluster's score; a representation of the performance of that cluster for the general case which was not affected by scale.

The statistical analysis detailed in in later sections of this paper principally pertains to the performance of these clusters.

## Hyperparameters Considered

The randomised hyperparameter allocation algorithm selected, for a given model subspecies, relevant hyperparameters within a range of values. The range of each hyperparameter was set based on figures found throughout the literature.

Given the memory requirements associated with preparing data for training and testing, spatial and temporal lags were not randomised. Rather, an equal number of models with two, three, four, and five temporal lags and one and two spatial lags were developed and a forecast horizon of one temporal lag was produced by the models.

For the two model genera examined in this study, ANNs and SVMs, the randomised hyperparameters and the ranges of their values are detailed below.

### *Artificial Neural Network Hyperparameter Sets*

The hyperparameters investigated exclusively in the ANN hyperparameter allocation included the number of layers in the network, the number of neurons in each layer, the learning function, the neuron activation function, the initial learning rate, the function for the adjustment of that learning rate throughout the learning process, the momentum of the algorithm, Beta1, Beta2 and epsilon. The summary of the set of swept values for each hyperparameter can be seen in table 1.

*Table 1 Ranges of ANN-specific hyperparameters*

| **Hyperparameter** | **Set of values** |
|---|---|
| Number of layers | {1,2,3} |
| Number of neurons in each layer | {n, n/2, n/4} where n is the number dimensions in the input layer. |
| The learning function | {L-BFGS, SGD, Adam} |
| Neuron activation function | {Identity, logistic, hyperbolic tangent, sigmoid} |
| Initial learning rate | [9e1,1e-5] |
| Learning rate function | {Constant, Adaptive} |
| Momentum | [0,1] |
| Beta 1 | [0.8,1] |
| Beta 2 | [0.9,1] |
| Epsilon | [9e-7,1e-9] |

The algorithm developed was based around the Python scikit-learn package (Pedregosa, et al., 2011) whereby inappropriate hyperparameters, e.g. the use of Beta1 in a SGD ANN, were ignored. This facilitated randomisation of all hyperparameters for the ANN genus for each ANN model developed with inherent rule based removal of unnecessary hyperparameters.

The package default batch size of 200 the default tolerance, after which the model ceased to learn, of $10^{-4}$ and the default maximum iterations of 200 were all applied. The former two of these should have been randomised and it is a flaw of this study that the defaults remained in place for these hyperparameters. This was due to human error.

The maximum iterations of 200 was intentionally selected to reduce the total runtime of the algorithm given cost constraints.

*Support Vector Machine Hyperparameter Sets*

The hyperparameters investigated exclusively in the SVM hyperparameter allocation included the kernel type, the margin width, the error parameter, gamma and the degree of the polynomial kernels. The range of values for these hyperparameters can be seen in table 2.

*Table 2 Ranges of SVM-specific hyperparameters*

| **Hyperparameter** | **Set of values** |
|---|---|
| Kernel | {linear, polynomial, radial base function, sigmoid} |
| ½ Margin width ($\varepsilon$) | [9e1,1e-5] |
| Error parameter | [9e1,1e-5] |
| $\gamma$ | [9e1,1e-5] |
| Polynomial degrees | {2,3,4} |

Like the case of ANNs, the rules to remove inappropriate hyperparameters inherent to scikit-learn was utilised and all hyperparameters were randomised for SVMs within the algorithm.

The default tolerance for SVMs of $10^{-3}$ was used and a maximum number of iterations of 15000 was used for the SVMs. Again, the former should have been randomised but was not due to human error and the latter was applied to reduce runtime while allowing for enough models to converge.

# Results and Analysis

This section reviews the steps taken in the analysis of experimental results. It begins with an analysis of the randomly selected sample of roads examined so as so understand the nature of roads that this analysis is likely to be valid for.

As mentioned in the methodology, results are analysed independently of the subject of each model (i.e. the roads) through the implementation of k-means clustering with respect to relevant hyperparameter values. The median of the accuracies of the models within each cluster was recorded as the clusters accuracy to minimise the bias extreme values may place on the clusters recorded performance.

Throughout this analysis, many t-tests were undertaken. To reduce the probability of a type I error, a significance level of 0.01 was applied for all t-tests. For the sake of consistency, this was applied to all hypothesis tests throughout the study.

## Analysis of Roads.

To understand the nature of urban roads that these experiments examined, an analysis of the roads was undertaken. This road analysis is aimed to aid in the interpretation of subsequent results.

### *Road Length*

The random sample of roads from the Aarhus dataset had the following summary statistics for road length:

*Table 3 Road length statistics for the sample and complete dataset*

| Statistic | Sample Value | Aggregate Value |
|---|---|---|
| Mean | 1.28e3 m | 1.15e3 m |
| Variance | 5.48e5 | 1.18e6 |

In an assessment of the null hypothesis that the population means are equal by Welch's t-test, a variant of Student's t-test that does not require equal variance in the samples being compared, the null hypothesis was not rejected.

This result should be analysed following review of the skewed pseudo-normal distribution of road length in the sample. For the population value, such consideration is not required given the sample size (Ghasemi & Zahediasl, 2012).

.

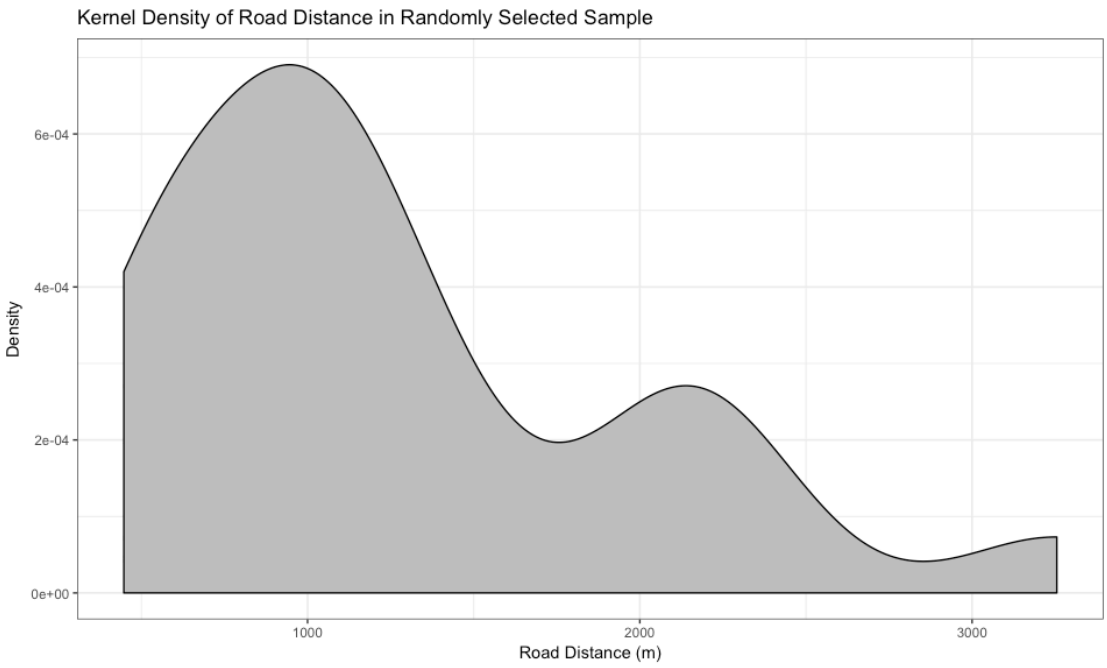



*Figure 3 A kernel density distribution of road lengths in the sample (Wickham, 2009).*

*Speed Limit*

The random sample of roads from the Aarhus dataset had the following summary statistics for speed limits:

*Table 4 Speed limit statistics for the sample and complete dataset*

| Statistic | Sample Value | Aggregate Value |
|---|---|---|
| Mean | 49.3 km/h | 44.8 km/h |
| Variance | 266 | 327 |

In an assessment of the null hypothesis that the population means of the speed limit are equal by Welch's t-test, the null hypothesis was not rejected. This result should be analysed following review of the skewed pseudo-normal distribution of speed limits in the sample. For the population value, such consideration is not required given the sample size (Ghasemi & Zahediasl, 2012).

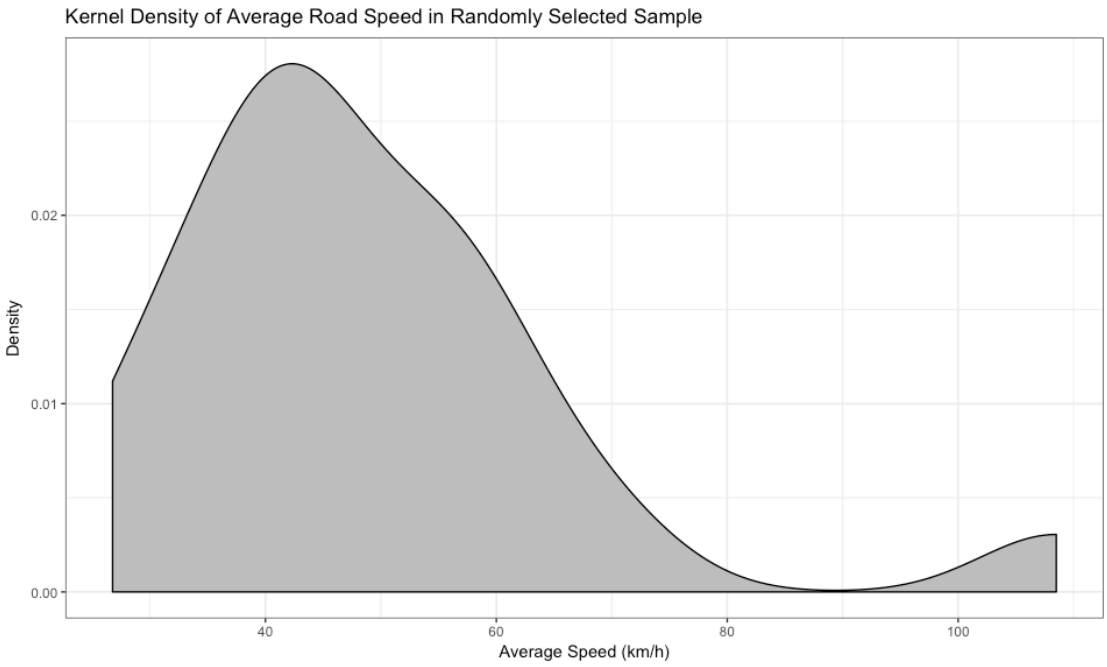



*Figure 4 A kernel density distribution of speed limits in the sample (Wickham, 2009).*

*Network Metrics*

Given the extent of the representative network was not exhaustive, analysis of the distribution of network-wide variables, e.g. connectedness, was not appropriate. As such, a combination of vertex and edge specific graph metrics, namely the degrees, was assessed and the test sample was compared to the population of roads in the dataset. In this context, the degree of a road is defined as the number of roads that it is immediately connected to.

The random sample of roads from the Aarhus dataset had the following summary statistics for degree:

*Table 5 Road degree statistics for the sample and complete dataset*

| Statistic | Sample Value | Aggregate Value |
|---|---|---|
| mean | 4.3 | 4.49 |
| variance | 4.75 | 8.41 |

In an assessment of the null hypothesis that the population means of degree are equal by Welch's t-test, the null hypothesis was not rejected. This result should be analysed following review of the skewed pseudo-normal distribution of degree in the sample. For the population value, such consideration is not required given the sample size (Ghasemi & Zahediasl, 2012).

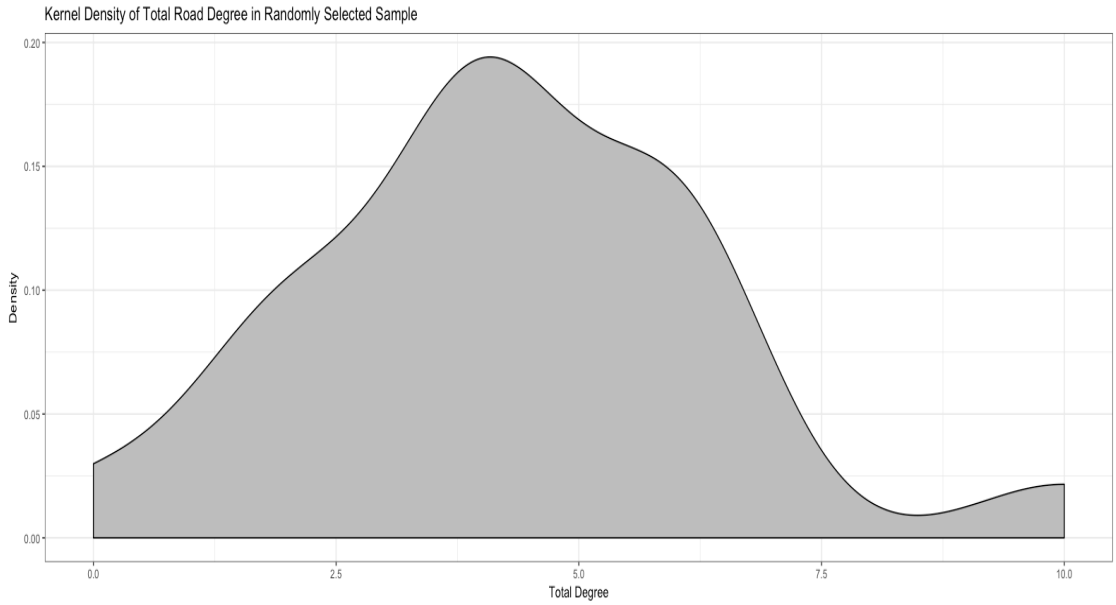



*Figure 5 A kernel density distribution of road degree in the sample (Wickham, 2009).*

### *Temporal Autocorrelation*

The extent of temporal correlation for the subset ranged from eight through to 18 temporal lags, as determined by examination of partial autocorrelation plots. These plots were highly periodic and showed insignificant correlation at one temporal lag frequently. Because of this, the minima for temporal lags was set to two. Given the number of the partial autocorrelation plots, it was not suitable to include them in the main text of this paper, however they can be found in Appendix A.

## Hyperparameter Analysis

In understanding the relationship between hyperparameters in both ANN and SVM road traffic forecasting models, several statistical tests were conducted.

This assessment pertained to both quantitative and qualitative hyperparameters, and as such, different tests were performed on these different data types. Comparison of genera, species and subspecies with different qualitative hyperparameters was

undertaken using Welch's t-test. Given the sufficient sample sizes, violation of the usual precondition for normality was not of concern (Ghasemi & Zahediasl, 2012) and was not applied as a precondition for those tests.

The initial null hypothesis tested for different sub-samples, as marked by qualitative hyperparameters was that the population means of the two samples being compared was equal, i.e.;

$$H_{0_1}: \mu_i = \mu_j$$

This was calculated using the two-tailed Welch's t-test with the absolute value of the differences of sample means being used.

If this null hypothesis was rejected, comparing the sample means and halving of the probability, converting the test to a one-tailed test facilitated analysis of a second null hypothesis;

$$H_{0_2}: \mu_i \preccurlyeq \mu_j$$

Where *i* was the sample with the lower sample mean and *j* was the sample with the higher sample mean.

The above hypotheses were tested for all combinations of qualitative hyperparameter values at the appropriate level of model classification specificity. For ease of digestion, readers will be presented with the resultant hypotheses, $H_R$, which may include the null

hypotheses or alternative hypotheses. Where possible, these hypotheses were condensed into a single line expression. For full results of all t-test conducted, see Appendix B.

For the analysis of quantitative hyperparameters, the plots of each hyperparameter against each error metric were examined at appropriate levels of model classification specificity. In the case that the plot indicated a relationship may exist, appropriate statistical tests were undertaken.

Given that a pseudo-linear relationship was not present for more than one hyperparameter for any genera, species or subspecies, multiple linear regression was not appropriate in this study (Williams, et al., 2013).

However, the possibility that the multidimensional hyperparameter space masked trends for hyperparameters remained. To unmask the relationship between individual hyperparameters and error, plots of the error minima for a given bin width of hyperparameter values against the mean of the bin's hyperparameter values inspected. This provided a means of investigation of trends toward error minima.

This was akin to understanding trends in the hyperparameters when all other quantitative hyperparameters were near optimised. Given the extent of this masking, all plots throughout this section are binned minima plots unless specified otherwise.

Prior to analysis of clustered results, clusters containing the most extreme value for error, of 1000, were removed from consideration. This was an upper bound for the

accuracy metric and as such, the accuracy of these clusters was likely greater than 1000, and not equal to 1000 meaning unreliable accuracy was recorded.

*Genera Analysis*

In comparing the relative accuracy of SVM and ANN models within the range of hyperparameters examined, the null hypotheses of equality population means was rejected for both MAE and MEDAE.

In testing the second null hypothesis, that the mean of the ANN population was less than the mean of the SVM population, the evidence was sufficient to reject this second null hypothesis and adopt the resultant hypothesis;

$$H_R: \mu_{ANN} < \mu_{SVM}$$

For the case of spatial lags for all models, the null hypothesis of equality between spatial and aspatial models was rejected for both MAE and MEDAE and subsequent testing of the second null hypothesis resulted in the adoption of the following resultant hypothesis for both error metrics:

$$H_R: \mu_{spatial} < \mu_{aspatial}$$

For the case of temporal lags for all models when the error was measured by MAE, the following resultant hypotheses were adopted:

$H_R$:

$\mu_2 = \mu_3$

$\mu_2 < \mu_4$

$\mu_2 = \mu_5$

$\mu_3 < \mu_4$

$\mu_3 = \mu_5$

$\mu_4 = \mu_5$

For the case of temporal lags when the error was measured by MEDAE, the following resultant hypotheses were adopted:

$H_R$:

$\mu_2 = \mu_3$

$\mu_2 < \mu_4$

$\mu_2 < \mu_5$

$\mu_3 = \mu_4$

$\mu_3 = \mu_5$

$\mu_4 = \mu_5$

For the ANN genus, the following resultant hypothesis pertaining to spatial lags was adopted for both error metrics.

$$H_R: \mu_{aspatial} < \mu_{spatial}$$

Additionally, the following resultant hypothesis pertaining to temporal lags was adopted for both error metrics.

$$H_R:$$

$$\mu_2 = \mu_3$$

$$\mu_2 = \mu_4$$

$$\mu_2 > \mu_5$$

$$\mu_3 = \mu_4$$

$$\mu_3 = \mu_5$$

$$\mu_4 = \mu_5$$

Note that these hypotheses are inconsistent with the general model case.

For the SVM genus in both the spatial and temporal hypothesis tests, none of the null hypotheses were rejected.

*Species Analysis*

This section is focused on the comparative analysis of species and sub-species within model genera. ANN models are first examined, followed by SVM models.

*Artificial Neural Networks*

**Species Comparison**

In assessing the marginal benefit of learning functions over one another, a series of Welch's t-tests were conducted. Following interpretation of two and, where appropriate, subsequent one tailed tests, the following resultant hypothesis was adopted.

$$H_R: \mu_{lbfgs} < \mu_{sgd} < \mu_{adam}$$

**ANN: L-BFGS**

In the species analysis of ANNs that optimise synapse weights through the L-BFGS algorithm, performance of the species with different activation functions was assessed. This began with an assessment of the null hypothesis that the mean performance of all activation functions was equal within this domain.

In the comparison of performance of activation functions, the following resultant hypothesis was adopted.

$$H_R:$$

$$\mu_{identity} = \mu_{logistic}$$

$$\mu_{identity} = \mu_{relu}$$

$$\mu_{identity} = \mu_{tanh}$$

$$\mu_{logistic} = \mu_{relu}$$

$$\mu_{logistic} = \mu_{tanh}$$

$$\mu_{tanh} < \mu_{relu}$$

In assessing the temporal lags within this species, the following resultant hypothesis was adopted.

$$H_R:$$

$$\mu_2 = \mu_3$$

$$\mu_2 > \mu_4$$

$$\mu_2 = \mu_5$$

$$\mu_3 > \mu_4$$

$$\mu_3 = \mu_5$$

$$\mu_4 = \mu_5$$

In assessing the spatial lags within this species no null hypotheses were rejected. Relationships between quantitative hyperparameters and model accuracy were assessed at the subspecies level.

*Identity*

For the ANN within this species with identity activation functions, all quantitative hyperparameter plots showed no aggregate relationship between error and the hyperparameters. Of note though is that there were striations in these plot, such as those seen in the non-minimised plot of the number of model synapses and MAE below. This may suggest that this species tends towards levels of accuracy following its optimisation process. This was also the case for logistic and rectified linear activation functions.

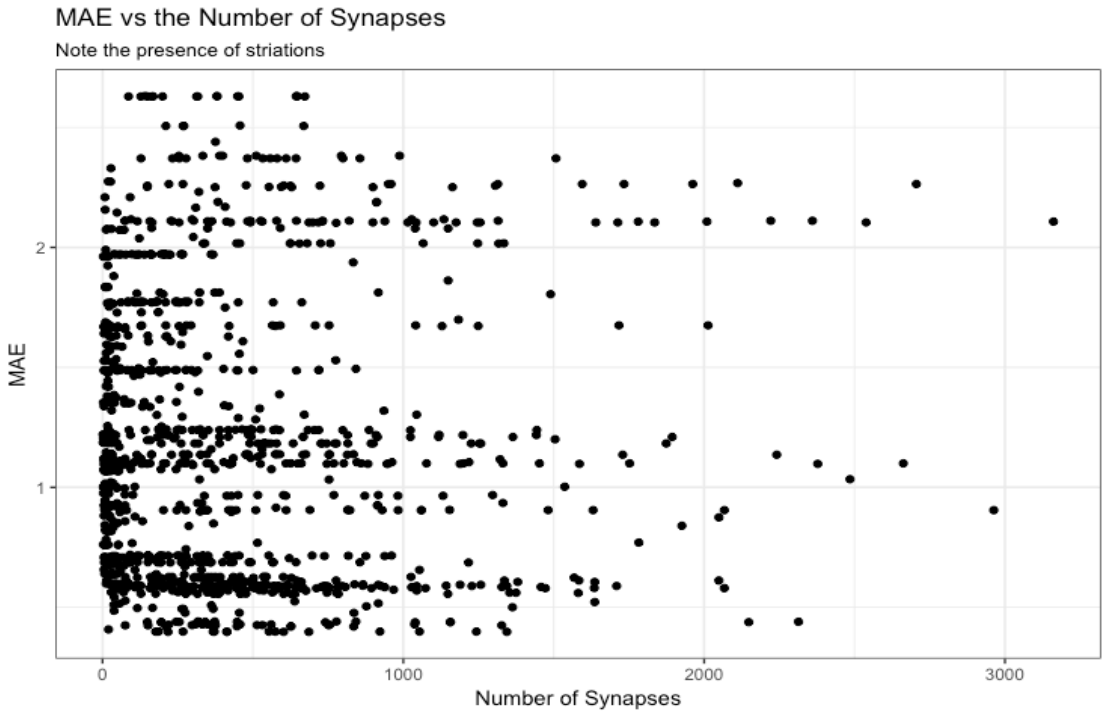



*Figure 6 A plot showing striations in model accuracy for an L-BFGS/Identity ANN (Wickham, 2009).*

*Logistic*

All quantitative hyperparameter plots showed no aggregate relationship between error and the hyperparameters.

*Rectified Linear*

All quantitative hyperparameter plots showed no aggregate relationship between error and the hyperparameters.

*Hyperbolic Tangent*

All quantitative hyperparameter plots showed no aggregate relationship between error and the hyperparameters.

In the minima plots of model complexity, as measured by total synapses, a linear relationship was seen to exist, however the heteroscedasticity of the relationship rendered linear regression inappropriate.

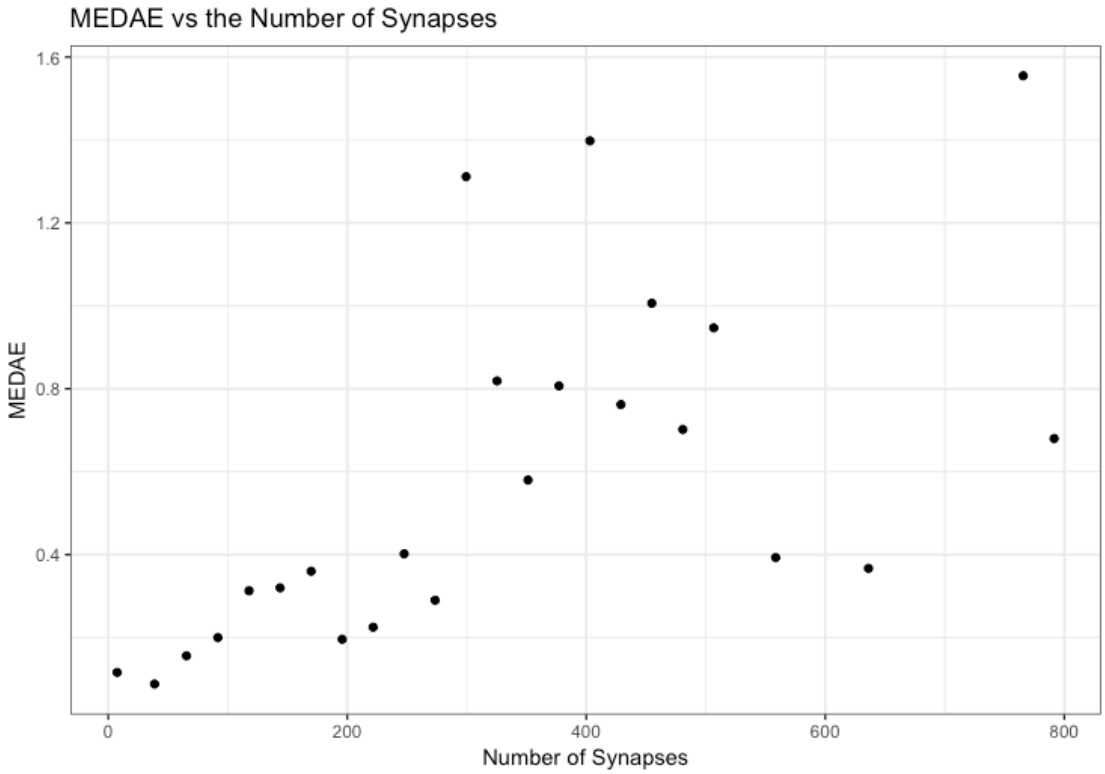



*Figure 7 A pseudo-linear relationship between MEDAE and network complexity, as measured by the total number of synapses (Wickham, 2009).*

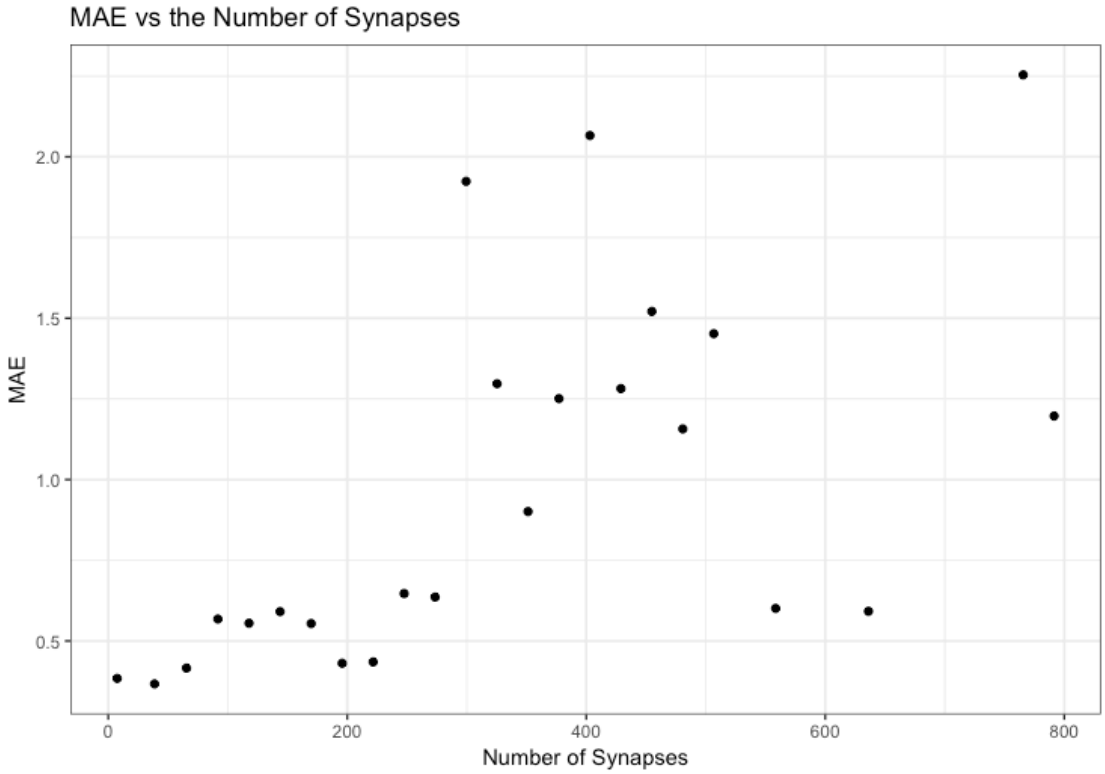



*Figure 8 A pseudo-linear relationship between MAE and network complexity, as measured by the total number of synapses (Wickham, 2009).*

**ANN: SGD**

For the case of SGD ANNs, prior to classification of the sub-species at the level of activation function, the effect of different learning rate functions was assessed. The two functions assessed were constant and adaptive. The null hypothesis that the population mean performance of these two populations was equal was not rejected for any error measurement. As such, SGD's were assessed in aggregate.

In assessing the relative performance of different activation functions, the following resultant hypotheses were adopted.

$$H_R:$$

$$\mu_{identity} < \mu_{logistic}$$

$$\mu_{identity} = \mu_{relu}$$

$$\mu_{identity} = \mu_{tanh}$$

$$\mu_{logistic} = \mu_{relu}$$

$$\mu_{logistic} = \mu_{tanh}$$

$$\mu_{tanh} = \mu_{relu}$$

In assessing both spatial and temporal lags, no null hypotheses were rejected. Relationships between quantitative hyperparameters and model accuracy were assessed at the subspecies level.

*Identity*

For the ANNs within this species with identity activation functions, there were low-error records throughout the spectrum of values for all quantitative hyperparameters.

Accounting for scale in error, the relationship between momentum and MEDAE appears to be quadratic, as seen in figure 9. With the probability that this relationship was random being 1.05e-3, the hypothesis that a quadratic relationship exists was adopted for MEDAE.

The heteroscedasticity of the data about a quadratic locus with respect to MAE was too extreme and as such, the test conducted for MEDAE was inappropriate.

*Logistic*

For ANNs within this species with logistic activation functions, there were low-error records throughout the spectrum of values for all quantitative hyperparameters.

In many of the plots, a small set of outliers existed. One potential reason for this may be that the combination of hyperparameters increased the number of iterations required for the ANN to converge, and, for the maximum iterations set in the random hyperparameter allocation algorithm, these ANNs were unable to converge. This was also the case for the rectified linear and hyperbolic tangent activation functions.

In both the minima and regular plots, there were no observed relationships.

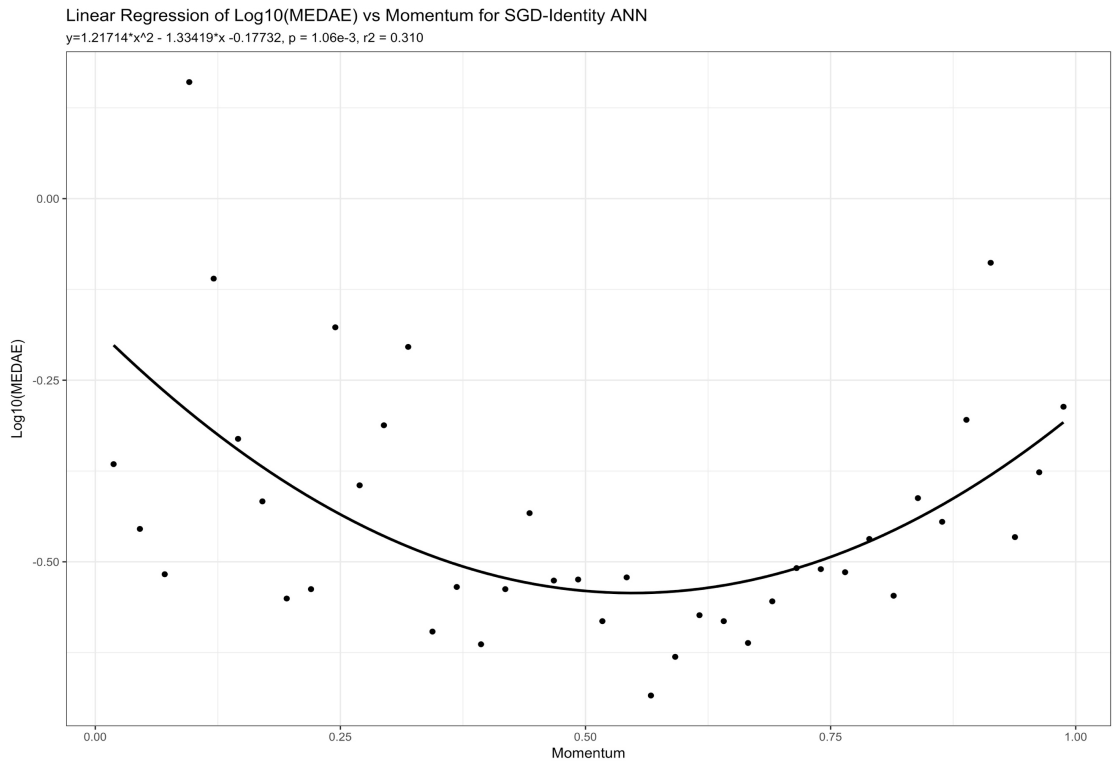



*Figure 9 The quadratic relationship between momentum and the logarithm of MEDAE for SGD/Identity ANNs (Wickham, 2009).*

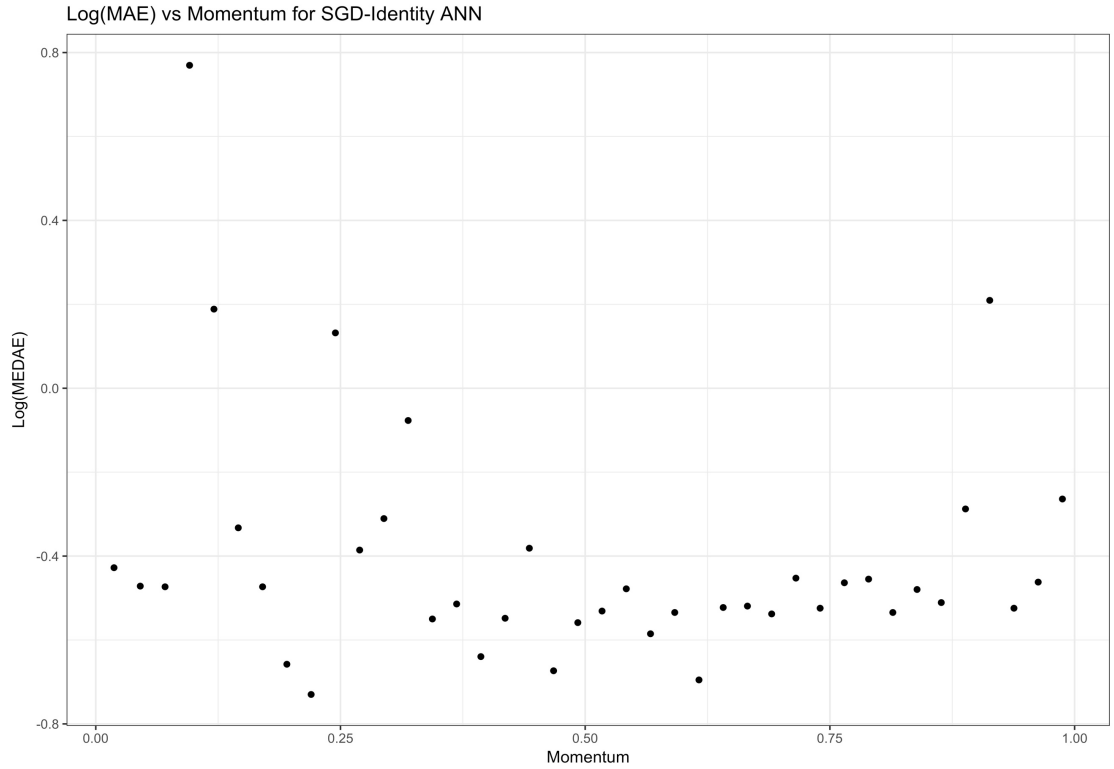



*Figure 10 The pseudo-quadratic relationship between momentum and the logarithm of MAE for SGD/Identity Linear ANNs (Wickham, 2009).*

*Rectified Linear*

For ANNs within this species with rectified linear activation functions, there were low-error records throughout the spectrum of values for all quantitative hyperparameters.

Reviewing the binned log-log minima plot for initial learning rate, excusing outliers, the hypothesis that no linear relationship exists between initial learning rate and model accuracy came into question. With the probability that a linear relation was random of 3.39e-4 and 2.61e-3 for MEDAE and MAE respectively, the hypothesis that a linear relationship in is space exists was adopted.

With $r^2$ values of 0.816 and 0.660 for MEDAE and MAE respectively, the variation in model performance is well explained by the initial learning rate. Given that this is within the log-log space, the inference was made that a power law exists between initial learning rate and model accuracy.

The short extent of the linear relationship within this feature space, however, reduces the reliability of this hypothesis.

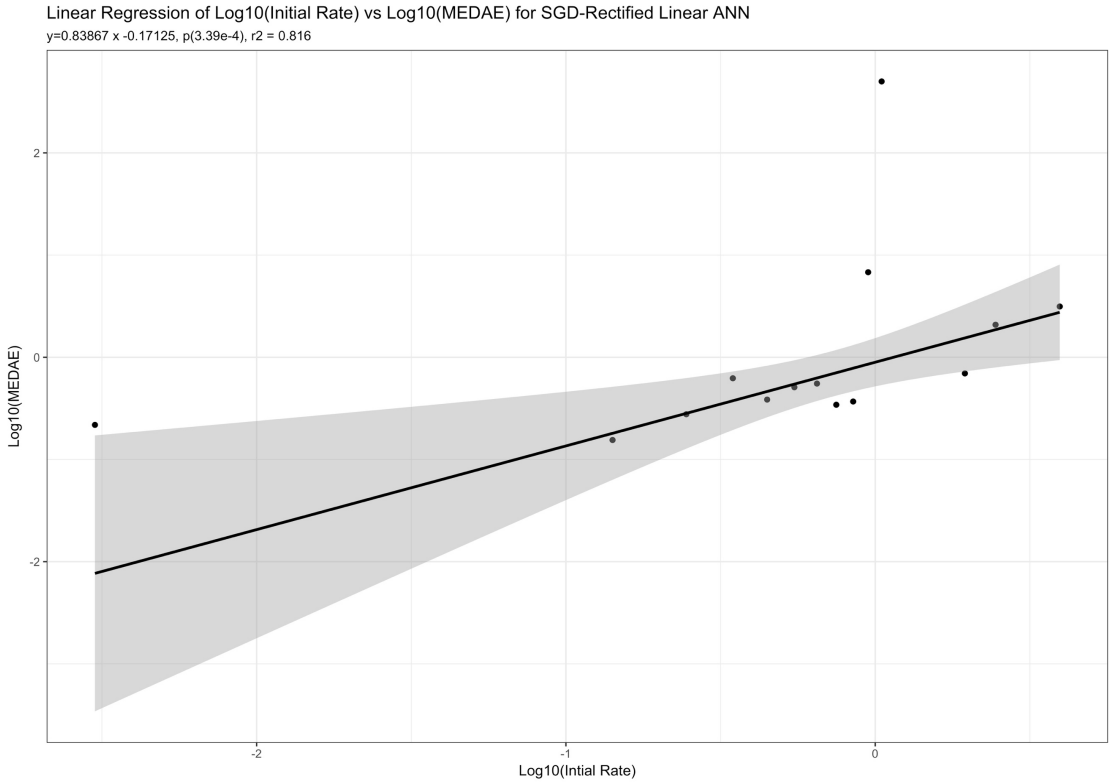



*Figure 11 The power law relation between initial learning rate and MEDAE for SGD/Rectified Linear ANNs (Wickham, 2009).*

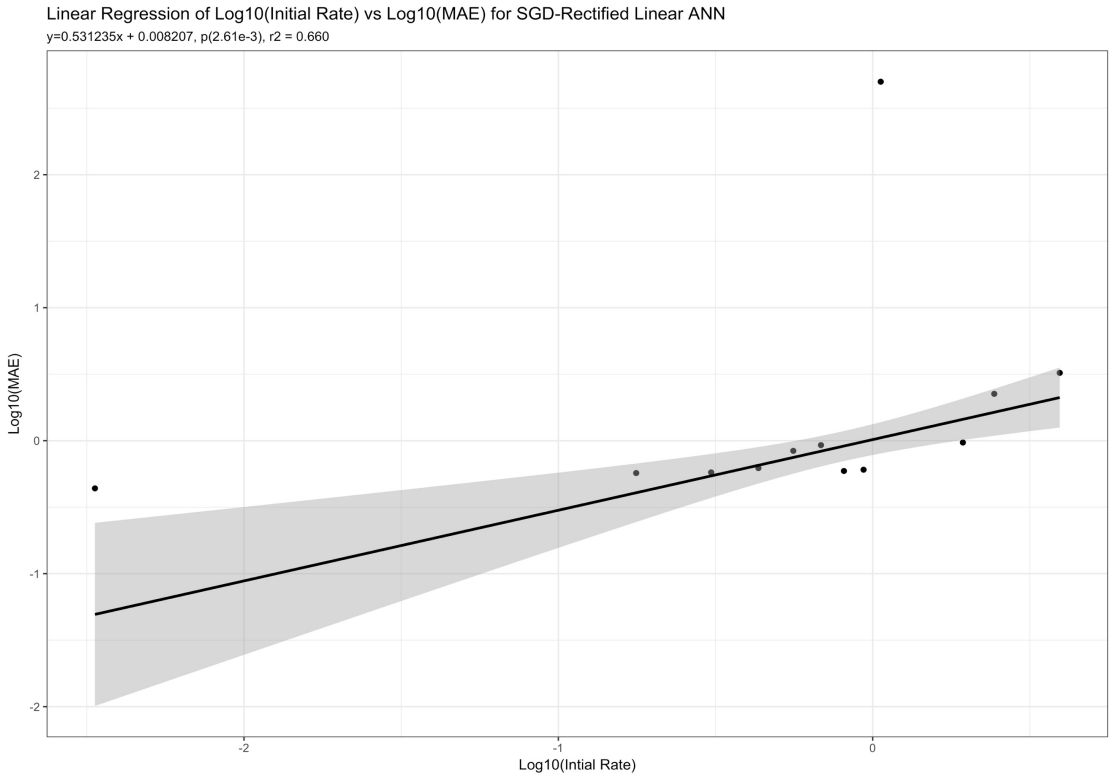



*Figure 12 The power law relation between initial learning rate and MAE for SGD/Rectified Linear ANN (Wickham, 2009)s.*

*Hyperbolic Tangent*

For the ANN within this species with hyperbolic tangent activation functions, there were low-error records throughout the spectrum of values for all quantitative hyperparameters.

In review of the minima plot for model complexity as measured by number of synapses, a pseudo-exponential function was seen. Upon review of the log-normal plot, the heteroscedasticity of data about an exponential locus rendered exponential linear regression in that space inappropriate and as such, the null hypothesis was not rejected.

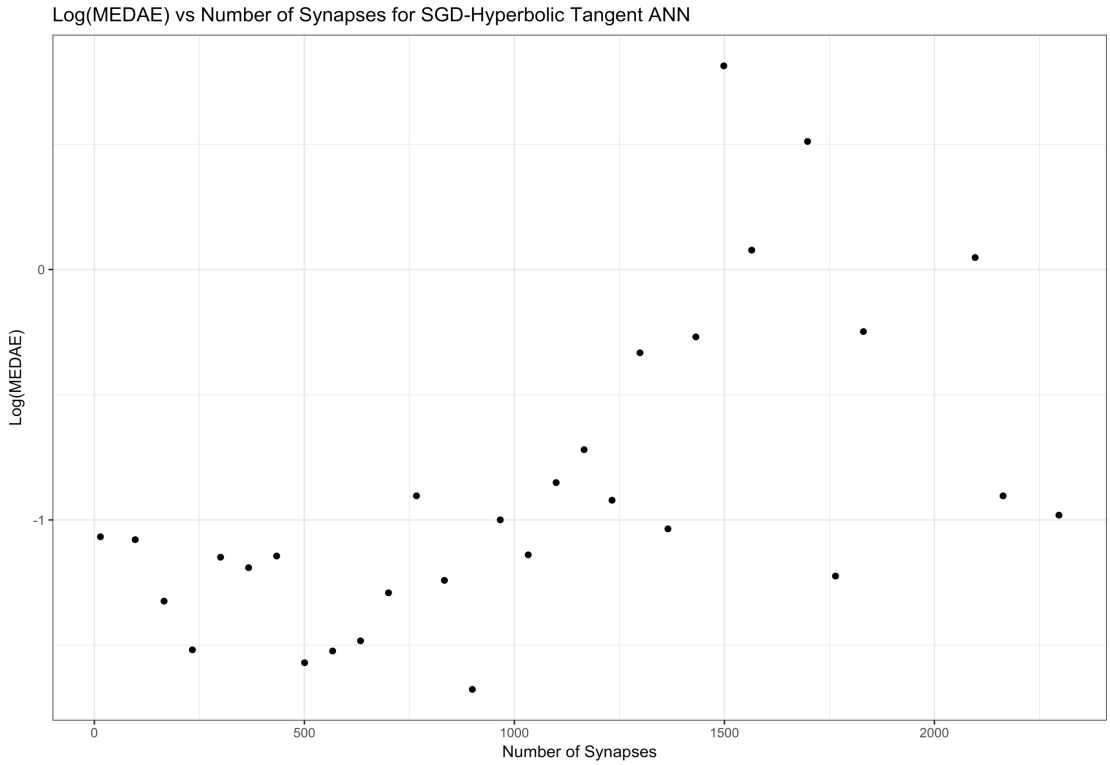



*Figure 13 The pseudo-exponential relationship between model complexity as measured by the number of synapses against MEDAE for SGD/Hyperbolic Tangent ANNs. In this plot, it is represented as a pseudo-linear relation (Wickham, 2009).*

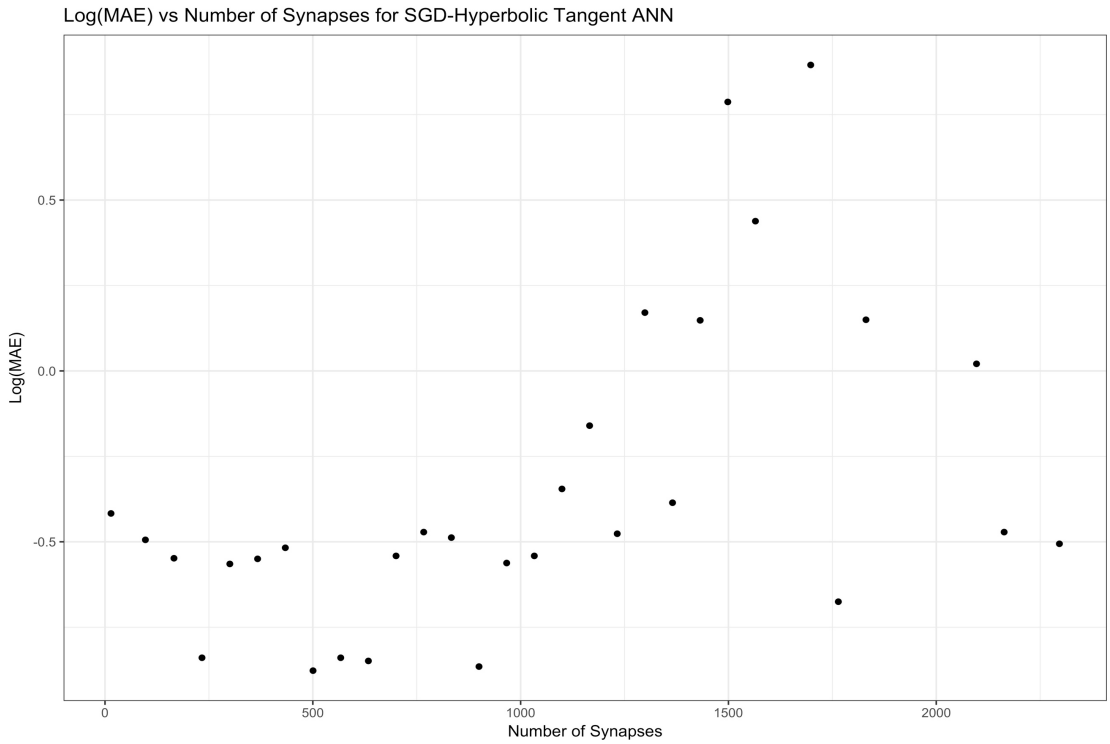



*Figure 14 The pseudo-exponential relationship between model complexity as measured by the number of synapses against MAE for SGD/Hyperbolic Tangent ANNs. In this plot, it is represented as a pseudo-linear relation (Wickham, 2009).*

**ANN: Adam**

In the species analysis of ANNs that optimise synapse weights through the Adam algorithm, performance of the species with different activation functions was assessed. This began with an assessment of the null hypothesis that the mean error of all activation functions was equal within this domain. Subsequently, the following resultant hypotheses were adopted for the respective error metrics.

MAE

$$H_R: \mu_{logistic} < \mu_{tanh} < \mu_{relu} < \mu_{identity}$$

MEDAE

$$H_R: \mu_{logistic} < \mu_{tanh} = \mu_{relu} < \mu_{identity}$$

In assessing the spatial and temporal lags, none of the null hypotheses were rejected.

*Identity*

For the ANNs within this species with identity activation functions, the hypothesis that a positive linear relationship between the model complexity, as measured by the total number of synapses and model error was tested.

Following the removal of outliers, the probability that a linear relationship between model complexity and error was random for MEDAE and MAE was found to be 2.65e-4 and 1.38e-4 respectively. As such, the hypothesis that a linear relationship exists between model accuracy and model complexity was adopted.

However, with $r^2$ values of 0.373 and 0.389 for MEDAE and MAE respectively, the variance in error explained by variance in model complexity is low.

There are two potential explanations for this. The first is that this is evidence of overfitting in complex models and the lack of more general predictive power therein. The second is that, given the iteration limit of 200, the more complex ANNs were not able to converge, thus affecting their accuracy.

*Logistic*

For the ANN within this species with logistic activation functions, the hypothesis that a positive linear relationship exists between the model complexity, as measured by the total number of synapses and model error was tested.

Following the removal of outliers, the probability that a linear relationship between model complexity and error was random for MEDAE and MAE was found to be 1.72e-7 and 2.06e-4 respectively. As such, the hypothesis that a linear relationship exists between model accuracy and model complexity was adopted.

An $r^2$ value of 0.629 for MEDAE suggests that the variance in error is explained moderately well by variance in model complexity. Conversely, an $r^2$ value of 0.363 for MAE suggests the variance in error explained by variance model complexity is moderately low.

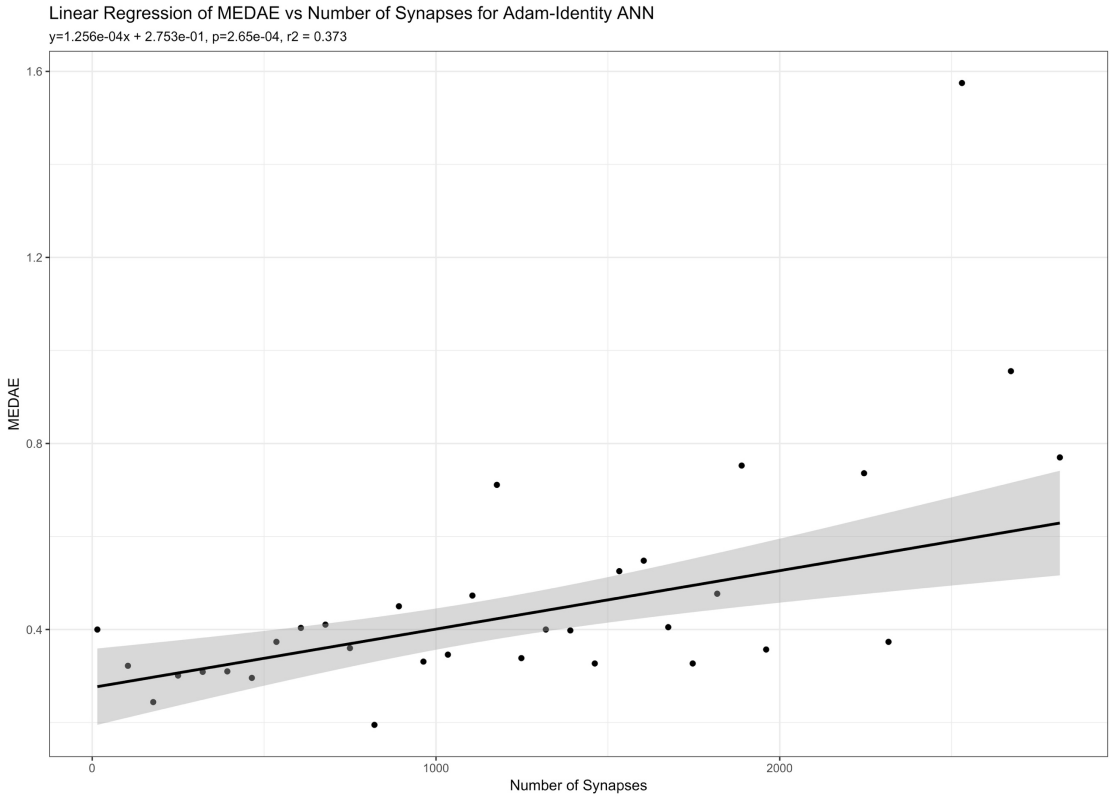



*Figure 15 The linear relationship between model complexity, as measured by the total number of synapses, and MEDAE for Adam/Identity ANNs (Wickham, 2009)*

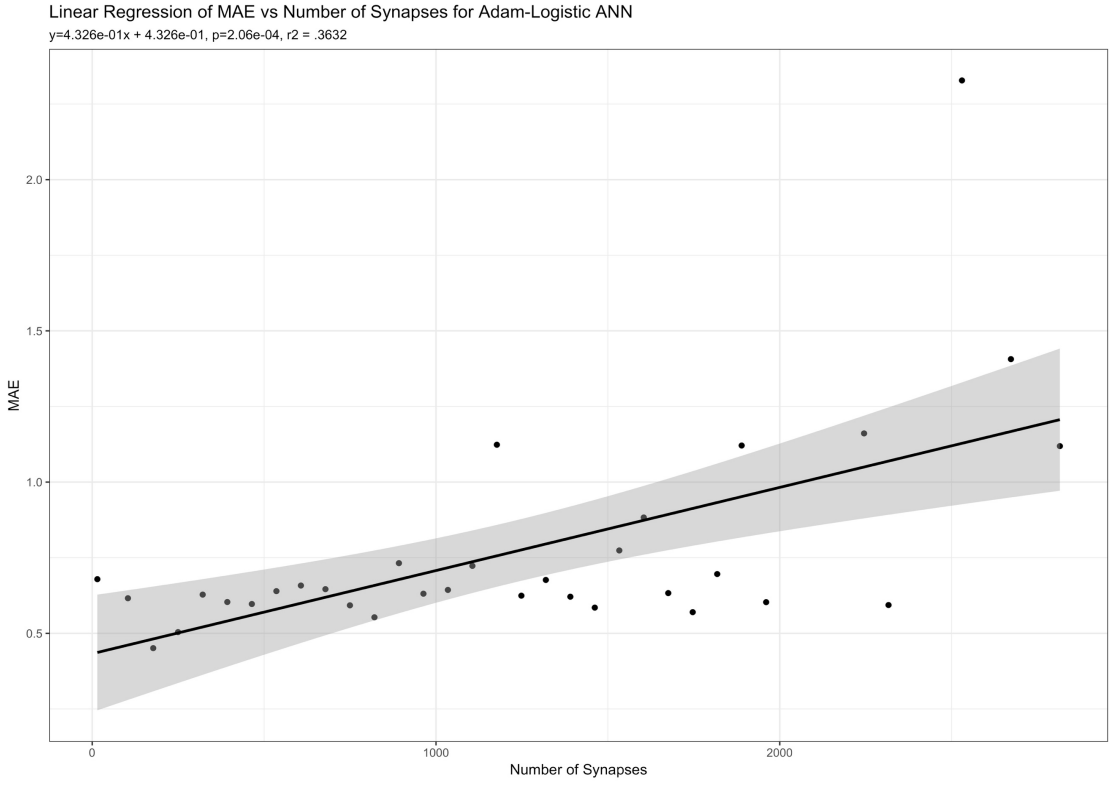



*Figure 16 The linear relationship between model complexity, as measured by the total number of synapses, and MAE for Adam/Identity ANNs (Wickham, 2009).*

Like the case for the identity activation functions in this species, there are two potential explanations for this. The first is that this is evidence of overfitting in complex models and the lack of more general predictive power therein. The second is that, given the iteration limit of 200, the more complex ANNs were not able to converge, thus affecting their accuracy.

Additionally, the plots in figure 17 and 18, show a mild heteroscedasticity about the fitted lines when the number of synapses exceeds 1500. As such, the reliability of these relationships in those regions is reduced.

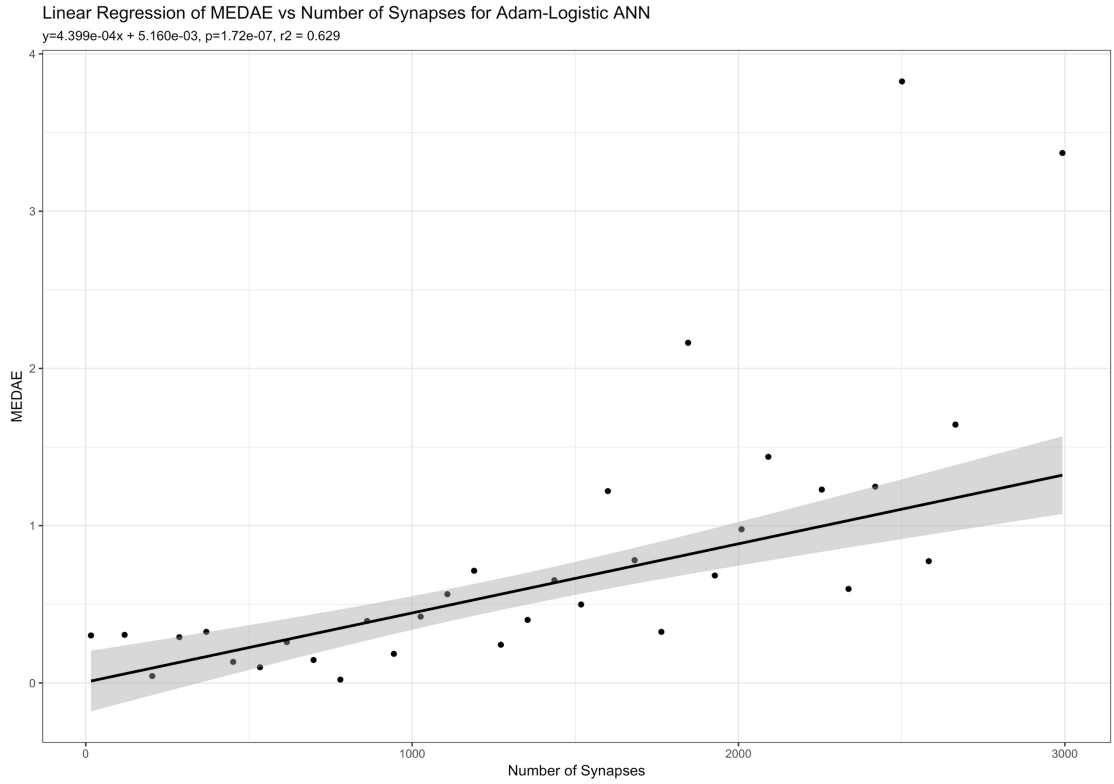



*Figure 17 The linear relationship between model complexity, as measured by the total number of synapses, and MEDAE for Adam/Logistic ANNs (Wickham, 2009).*

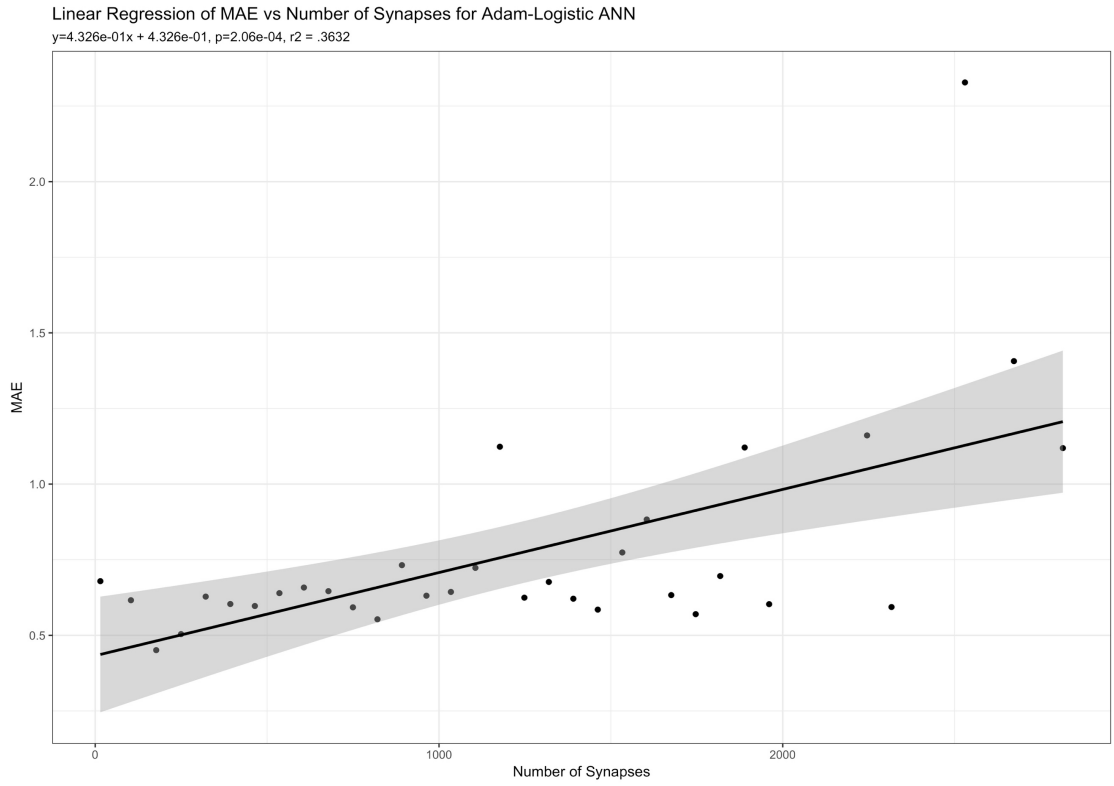



*Figure 18 The linear relationship between model complexity, as measured by the total number of synapses, and MAE for Adam/Logistic ANNs (Wickham, 2009).*

*Rectified Linear*

For the ANNs within this species with rectified linear activation functions, the hypothesis that an exponential relationship between the model complexity, as measured by the total number of synapses and model error was again tested.

Following the removal of outliers, the probability that an exponential relationship between model complexity and error was random for MEDAE and MAE was found to be 1.59e-6 and 6.43e-4 respectively. As such, the hypothesis that an exponential relationship exists between model accuracy and model complexity was adopted.

With $r^2$ values of 0.507 and 0.388 for MEDAE and MAE respectively, extent to which the variance in error is explained by variance in model complexity is moderate.

Like the case for the identity and rectified linear activation function in this species, there are two potential explanations for this. The first is that this is evidence of overfitting in complex models and the lack of more general predictive power therein. The second is that, given the iteration limit of 200, the more complex ANNs were not able to converge, thus affecting their accuracy.

Additionally, both plots show a mild heteroscedasticity about the fitted line as the number of synapses increases above 1500. As such, the reliability of this relationship within that region is reduced.

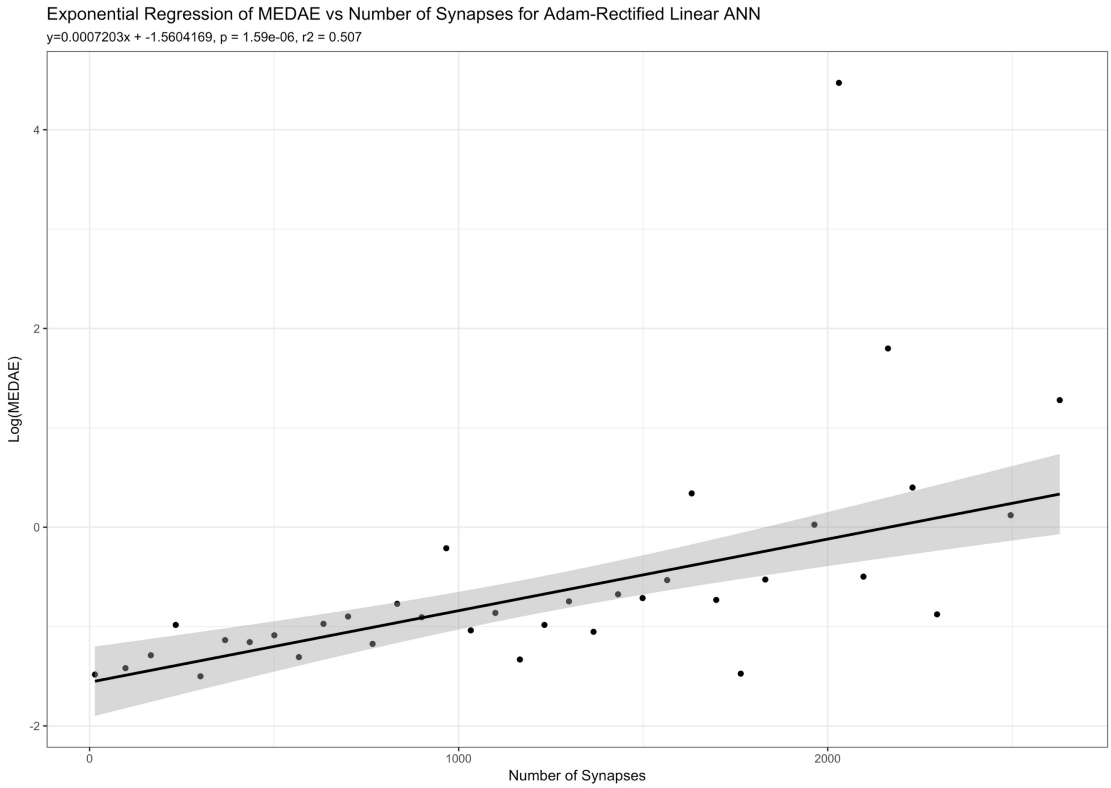



*Figure 19 The exponential relationship between model complexity, as measured by the total number of synapses, and MEDAE for Adam/Rectified Linear ANNs (Wickham, 2009).*

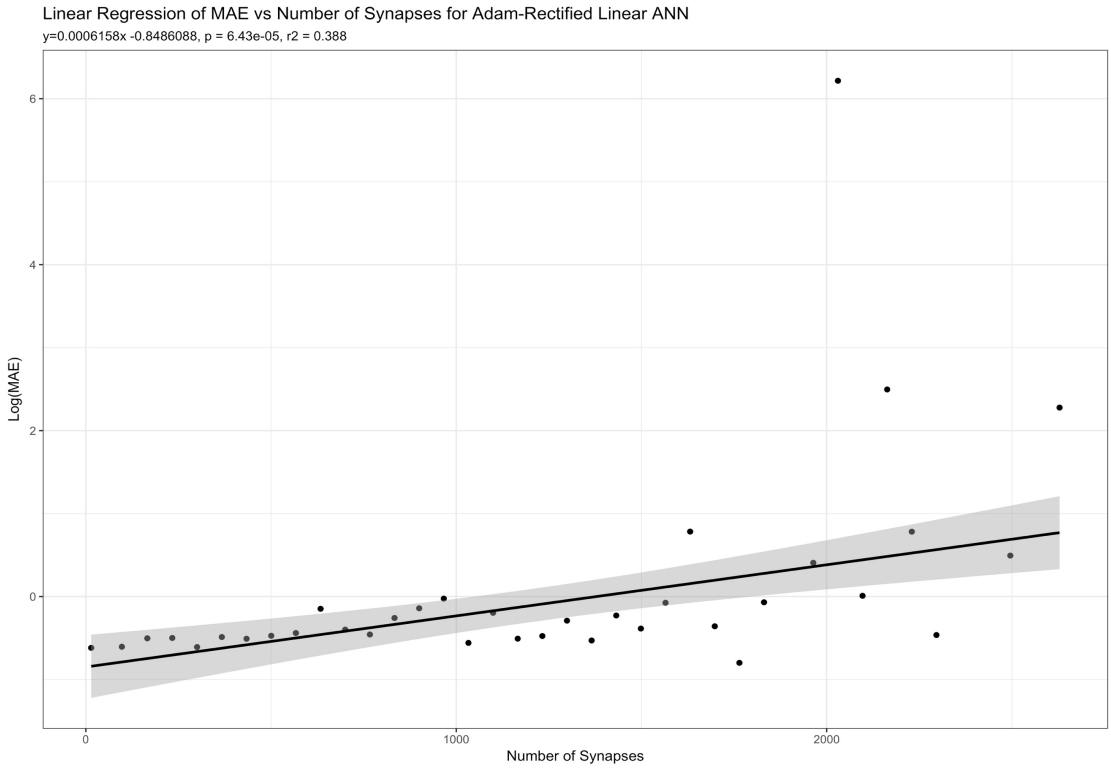



*Figure 20 The exponential relationship between model complexity, as measured by the total number of synapses, and MAE for Adam/Rectified Linear ANNs (Wickham, 2009).*

*Hyperbolic Tangent*

For the ANN within this species with hyperbolic activation functions, the hypothesis that a quadratic relationship exists between the model complexity, as measured by the total number of synapses and model error was tested.

Following the removal of outliers, the probability that a quadratic relationship between model complexity and error was random for MEDAE and MAE was found to be 1.78e-5 and 5.67e-7 respectively. As such, the hypothesis that a quadratic relationship exists between model accuracy and model complexity was adopted.

With $r^2$ values of 0.555 and 0.655 for MEDAE and MAE respectively, the variance in error is explained by variance model in complexity moderately well.

As is the case for the linear regressions for other activation functions within this species, there is a degree of heteroscedasticity whereby variance about the fitted line is greater for models where the number of synapses is greater than 1500. This reduces the reliability of the findings within this range.

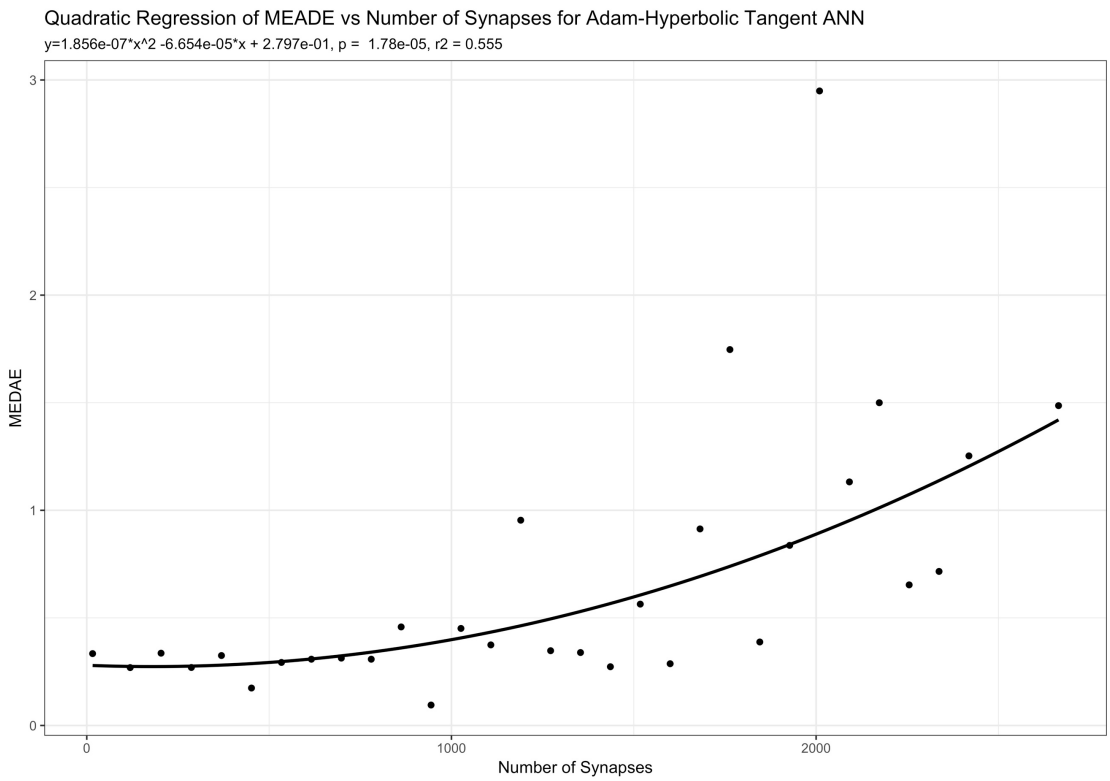



*Figure 21 The quadratic relationship between model complexity, as measured by the total number of synapses, and MEDAE for Adam/Hyperbolic Tangent ANNs (Wickham, 2009).*

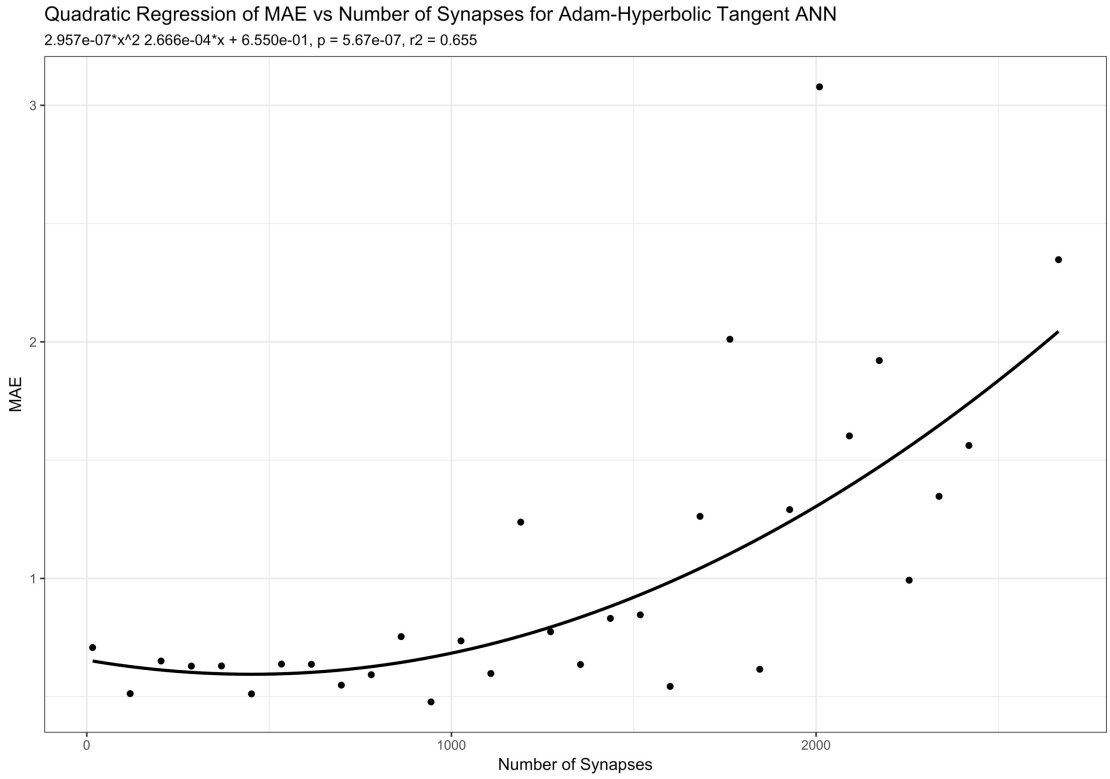



*Figure 22 The quadratic relationship between model complexity, as measured by the total number of synapses, and MAE for Adam/Hyperbolic Tangent ANNs (Wickham, 2009).*

*Support Vector Machines*

In assessing the marginal benefit of kernels over one another, a series of Welch's t-tests were conducted. Following interpretation of two and, where appropriate, subsequent one tailed tests, the following resultant hypotheses were adopted.

$$H_R:$$

$$\mu_{linear} < \mu_{polynomial}$$

$$\mu_{linear} < \mu_{rbf}$$

$$\mu_{ilinear} = \mu_{sigmoid}$$

$$\mu_{polynomial} = \mu_{rbf}$$

$$\mu_{polynomial} = \mu_{sigmoid}$$

$$\mu_{rbf} = \mu_{sigmoid}$$

**SVM: Linear**

For SVMs with a linear kernel a quadratic relationship was observed in the non-minima plot between the logarithm of half of the margin width, $\varepsilon$, and both MEDAE and MAE.

Following the removal of outliers, the probability that the quadratic relationship between model complexity and error was random for MEDAE and MAE was below the software's minimum bound of 2.2e-16 in both cases. As such, the hypothesis that a quadratic relationship exists between model accuracy and $\varepsilon$ was adopted.

With $r^2$ values of 0.568 and 0.533 for MEDAE and MAE respectively, the variance in error is moderately well explained by $\varepsilon$. The heteroscedacitiy of the data about the

fitted line is notably greater for $\varepsilon > 1$ and as such, the relationship is less reliable in this domain.

In assessing the spatial and temporal lags, no null hypotheses were rejected.

**SVM: Polynomial**

Classification of subspecies of model within the SVM Polynomial species may be made with respect to the degree of the polynomial kernel. In assessing the relative performance of different degrees of polynomial kernel function, a Welch's t-test was conducted, with the null hypothesis that all population means are equal. There was insufficient evidence to reject this null hypothesis and as such, the analysis of hyperparameters was conducted on the species in aggregate.

For this species, no relationships were found between hyperparameters and model error. Additionally, in assessing the spatial and temporal lags, no null hypotheses were rejected.

**SVM: Radial Base Function**

For SVMs with a radial base function kernel a sigmoid relationship was observed in the non-minima log-log plot between the logarithm of $\varepsilon$ against both MEDAE and MAE.

Following the removal of outliers, the probability that the sigmoid relationship between model complexity and error was random for MEDAE and MAE was below the software's minimum bound of 2.2e-16 in both cases. As such, the hypothesis that a sigmoid relationship exists between model accuracy and $\varepsilon$ was adopted.

With $r^2$ values of 0.871 and 0.868 for MEDAE and MAE respectively, the variance in error is explained well by this function of $\varepsilon$. In both cases, as $\varepsilon$ tends to zero, the variance about the sigmoidal locus increases. It is speculated that this may be a result of a lack of convergence for SVMs with low $\varepsilon$ values, which typically require a higher number of iterations to converge (Smola & Schölkopf, 2004). If these models had converged, the $r^2$ of the fit may have been higher and the explanatory power of $\varepsilon$ in this species accuracy would have increased.

In assessing the spatial lags, the following resultant hypothesis was adopted.

$$H_R: \mu_{spatial} < \mu_{aspatial}$$

In assessing the temporal lags, no null hypotheses were rejected.

**SVM: Sigmoid**

In assessing the spatial and temporal lags, no null hypotheses were rejected. Additionally, no relationships were observed in either the regular or the minima plots for any hyperparameters.

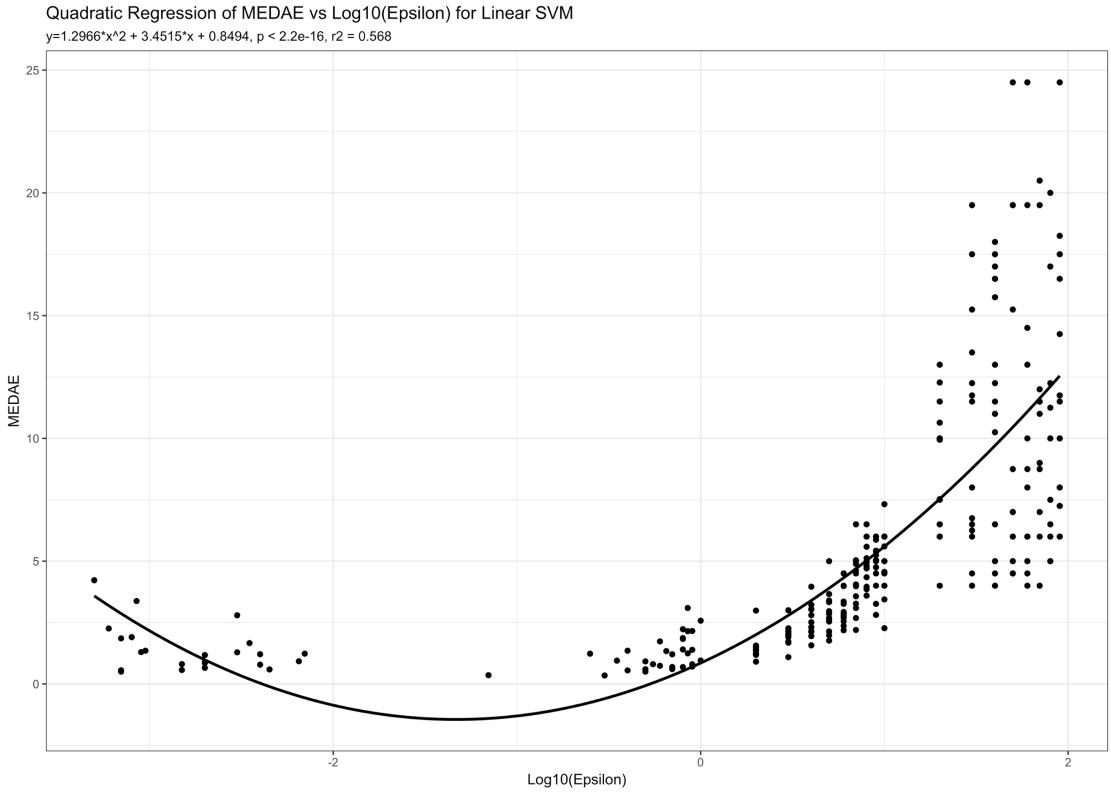



*Figure 23 The non-binned minima plot of the quadratic relationship between epsilon and MEDAE for Linear SVMs (Wickham, 2009).*

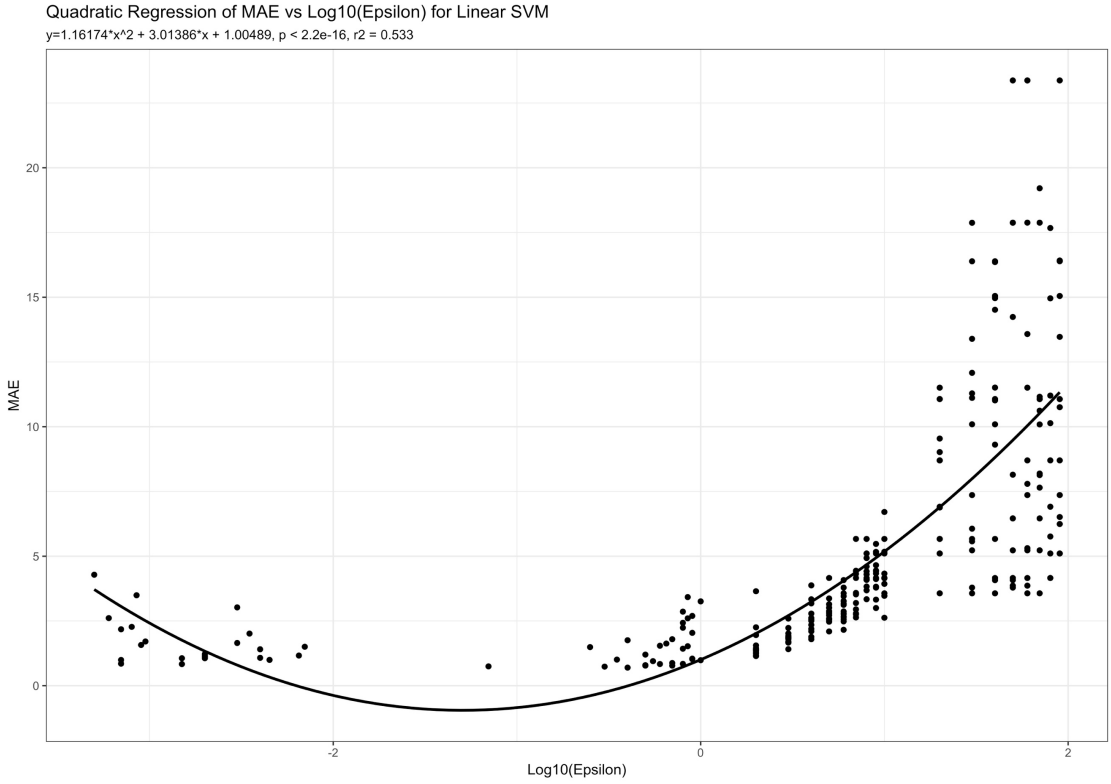



*Figure 24 The non-binned minima plot of the quadratic relationship between epsilon and MAE for Linear SVMs (Wickham, 2009).*

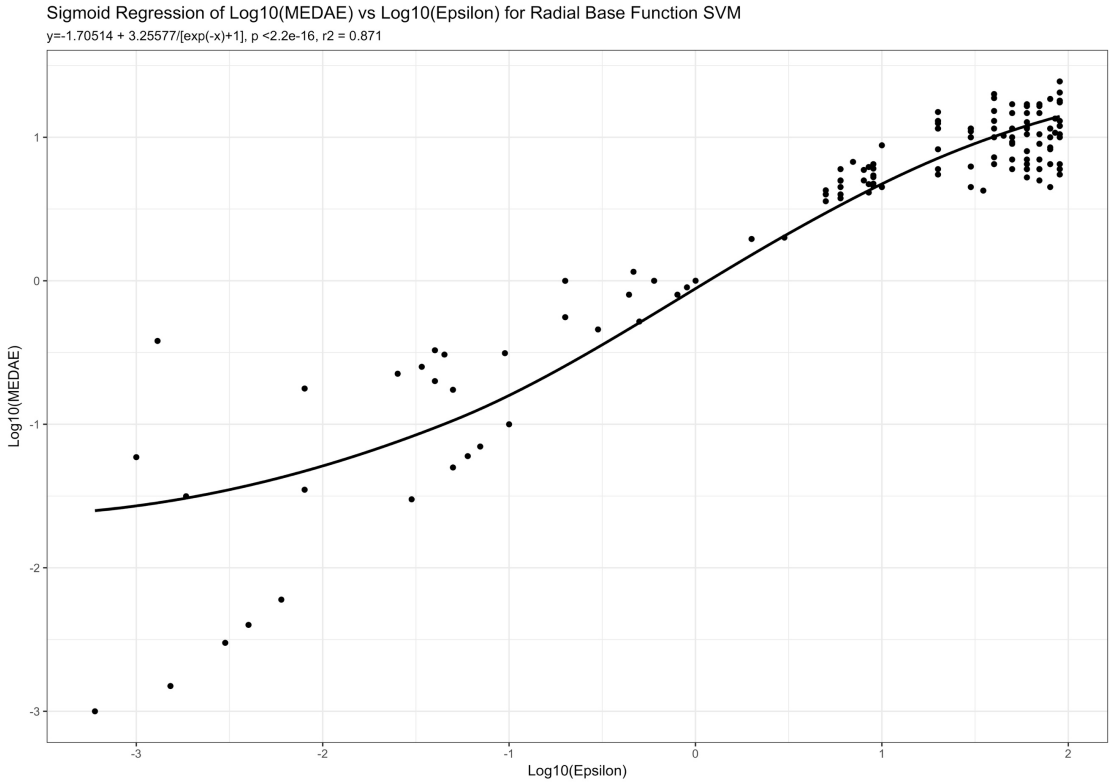



*Figure 25 The non-binned minima plot of the sigmoidal relationship between epsilon and MEDAE for Radial Base Function SVMs (Wickham, 2009).*

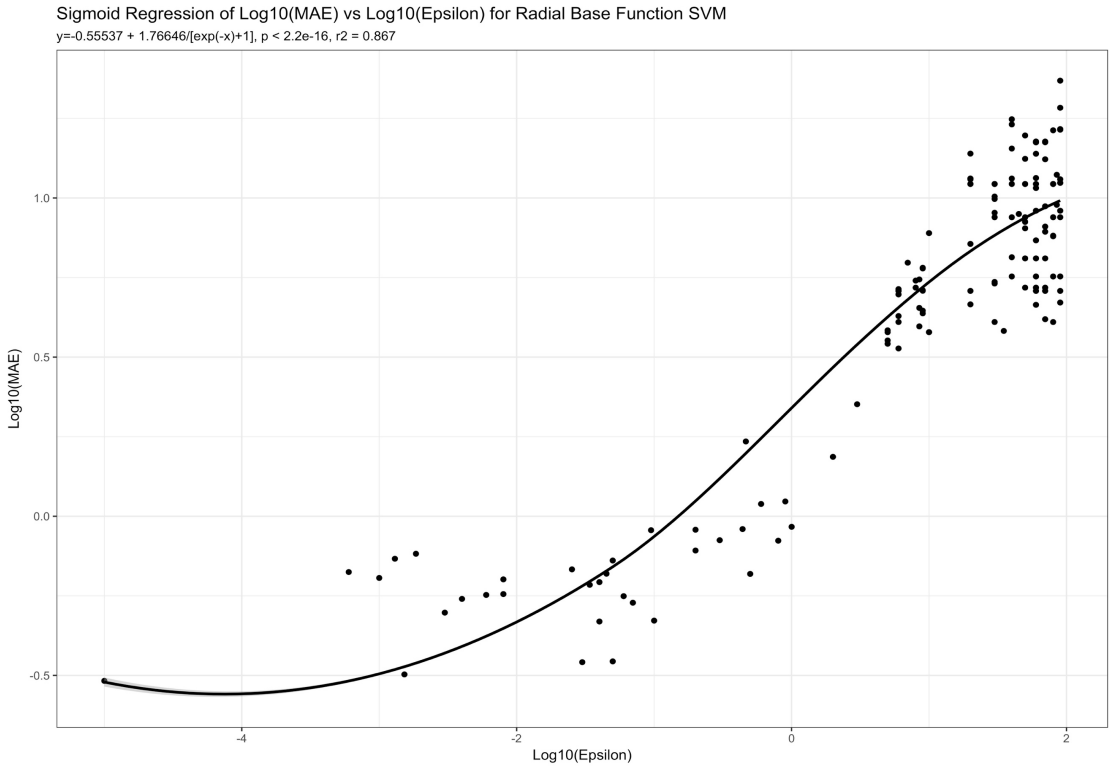



*Figure 26 The non-binned minima plot of the sigmoidal relationship between epsilon and MAE for Radial Base Function SVMs (Wickham, 2009).*

# Discussion

The aim of this paper is to contribute to the understanding of the nature of optimal road traffic forecasting models for the general case to better inform practitioners. As such, this discussion will focus on the reliability of the findings, the resultant hypotheses, and what those hypotheses mean for practitioners.

The methodology developed in this paper facilitated a robust statistical analysis of qualitative and quantitative hyperparameters in various model genera, species and subspecies through the implementation of a randomised hyperparameter allocation algorithm and the subsequent clustering of models based on their normalised, and relevant, hyperparameters.

The experimentation was conducted on a pseudo-randomly selected subset of 449 roads surrounding Aarhus, Denmark. In terms of road length, speed limit and degree, this sample was representative of the 449 roads from which they were selected.

The data was collected using an array of Bluetooth sensors (Bloksgaard & Christiansen, 2015). There were several caveats with using this data including the lack of full representation of minor roads, the recording of only vehicles with Bluetooth capability and reliability issues with the travel duration and speed data. The former two of these caveats were taken under consideration and the latter caveat was addressed through setting the forecasting metric as traffic flow.

Of note is that the reliability of the underlying dataset is inextricably linked to the reliability of the resulting findings. If the assumptions that the dataset fully represents the network and that the traffic flow is representative of the real flow are significantly flawed, the results found in this study may not generalise to other roads.

Practitioners should have an appreciation of the aforementioned caveats and the relevance they may have to their own forecasting context before accepting the findings of this study.

The temporal extent of the data examined was 13 weeks with a granularity of five minutes. Given the memory requirements associated with preparation of training and testing data, a fixed cross validation ratio of 3:1 was applied to all models. This requirement also necessitated limits to the combinations of temporal and spatial lags available in the allocation algorithm. Additionally, human error led to the omission of several hyperparameters from the allocation algorithm, including the tolerance of models and batch size. These compensations and errors limit the scope of the study within the wider hyperparameter feature space and hyperparameter values used outside of the scope may lead to different model performance.

With respect to the randomised hyperparameter allocation algorithm, the predetermined extent of the range of values from which the random number generators generated hyperparameters also limited the scope of this study within the hyperparameter feature space. However, following a literature review of applications of ANN and SVM models to traffic flow, the range of values is consistent with current best practice.

In analysing the resultant model data, the number of models developed prevented the inspection of atypical flows and the extent to which the models accurately forecasted such events. This is suboptimal given the desire of practitioners to forecast under these scenarios (Haworth, 2014). Additionally, analysis of such performance would not have been possible given the clustering approach taken to analyse hyperparameters without respect to the road upon which models were developed.

The development of an error metric to determine the atypical forecast accuracy of a model would be a valuable addition to the literature and would facilitate such analysis.

In the assessment of qualitative hyperparameters, two and one-tailed Welch's t-tests were applied to test null hypotheses of equality between relevant sub samples. Quantitative hyperparameter trends were reviewed in aggregate and in binned error minima plots, a tool developed in this study to unmask the trends when other hyperparameters were optimised.

Beginning with analysis of the spatiotemporal lags, models in aggregate were shown to perform better with the inclusion of a single spatial lag. This supports theses contenting that the inclusion of spatial data improves the accuracy of traffic forecasting models (Haworth, 2014) (Vlahogianni, et al., 2014). This result, however, was contradicted at finer levels of model classification specificity. The most consistent result, however, was that the population means for with and without spatial lags was equal.

This result may be interpreted as unreliable given the lack of network completeness for the dataset. Namely, the inclusion of spatial lags from a non-exhaustive representation of a network may not include all the neighbours of each road. As such, there may be little difference in inputs between models that include and do not include spatial lags. It is the advice of the author that, because of a lack of reliability in the network topology that spatial results in this paper be disregarded.

In the assessment of temporal lags, superior performance of less temporal lags was found at the aggregate level, and the inverse at some levels of higher model classification specificity, however the most common test results suggested little difference between the accuracy of models with different temporal lags.

This result may be interpreted as an error manifesting from insufficient temporal granularity. It may also be the case that higher temporal lags had little bearing on the final regression output in all cases, and perhaps contributed to model noise to inhibit model conversion. As such, repetition of this study with more temporally granular data may return different results.

Reviewing model genera, ANNs were found to outperform SVMs in aggregate for this dataset.

Within the ANN genera, the L-BFGS learning algorithm was shown to outperform all others in the general case. This is speculated to be due to the second order nature of its optimisation which facilitates optimisation in less iterations than first order

processes. If the maximum number of iterations was higher, differences between the performance of learning algorithms may not have been as pronounced. Additionally, the Adam species was found to have the highest error, a finding that contradicts the single sample comparison for classification found in the establishing paper for the learning function (Kingma & Ba, 2015).

Within the L-BFGS species, the performance of activation functions suggested that all models performed equally except for the hyperbolic tangent and rectified linear functions, where the hyperbolic tangent function performed better.

For the Adam species, logistic functions outperformed all other activation functions, whereas for the SGD species, all performed equally except for the identity function which outperformed the logistic function.

For each sub-species within the L-BFGS species, there were no significant relationships observed. Of note is that error for the hyperbolic tangent activation function sub-species did appear to have a pseudo-linear relationship with model complexity with other hyperparameters optimised.

For the case of the SGD species, momentum and the logarithm of error showed a quadratic relationship in the identity activation function sub-species with a minimum error near a momentum of 0.55. A power law was shown to exist between error and the initial learning rate for SGD species with a rectified linear function when other hyperparameters were optimised. Linear relationships between network complexity and error were also common for SGD ANNs.

For the case of Adam species, for all activation functions, both linear and quadratic trends were found for error and model complexity.

For cases were linear or quadratic trends between model complexity and model error were found, either lack of convergence or overfitting is speculated as the underlying cause. To assess this further, an algorithm like that used in this study should be run with a higher number of maximum iterations.

Within the SVM genera, the linear kernel was shown to outperform all other kernels. For all quantitative hyperparameter relationships found within the SVM genera, the relationships were present in the non-binned plots; thus, the effect of additional hyperparameters had little effect on the relationships that were shown.

For the linear kernel, the relationship between epsilon and error was shown to be convex, such that a minimum was located near $\varepsilon$ = 1e-1. However, this trend was only shown for the MEDAE case.

The radial base function showed a strong sigmoidal relationship between epsilon and error in the log-log feature space. Practitioners using this kernel would be well advised to use epsilon values less than 1.0e-2 if they have the computational resources to provide the machine sufficient iterations to converge.

All other SVM species showed no relationships between quantitative hyperparameters and error.

Practitioners should note that, for the non-binned plots, model error was frequently low across a broad spectrum of quantitative hyperparameters. This supports the practice of model tuning as high model accuracy can be obtained for most hyperparameters values, however, this does not support the status quo of the literature.

If research for road traffic forecasting models is to be of benefit to practitioners, the results must be reliable. The status quo of the field, i.e. single or small sample comparison of sample statistics (Castro-Neto, et al., 2009) (Lv, et al., 2015) (Kingma & Ba, 2015) (Liu, et al., 2011) (Min & Wynter, 2011) (Moretti, et al., 2015) (Yasdi, 1999), is not sufficient and may mislead practitioners. Of interest is that this status quo pertains not just to applications of machine learning to traffic forecasting; it seems near ubiquitous throughout machine learning literature generally (Kingma & Ba, 2015).

This study sheds light onto the impact that the advent of cloud computing and open source machine learning programs may have on the field of traffic forecasting and machine learning generally.

Through thorough statistical analysis of hyperparameters of models trained on generally representative datasets, researchers may be able to continue to reliably identify relationships between hyperparameters and model accuracy. The subsequent publication of such results would provide practitioners with statistically reliable hyperparameter trends and may have far reaching effects.

The extent to which the results of this study might provide practitioners with reliable findings is dependent upon the extent to which the Aarhus dataset is representative of the general case.

In terms of broader analysis, if the Aarhus dataset is considered representative of traffic generally, the findings of this study may be interpreted to understand more about the nature of traffic itself.

If the cause of the linear relationship between model complexity and error repeatedly shown in the ANN subspecies is overfitting and not premature cessation of learning, sections of road might be shown to behave differently over time. That is, if non-linear models that can accurately model relationships on a training set cannot forecast traffic flow on the same road at a different time, the dynamics of the road can be thought to have changed.

The extent to which such relationships are temporally dynamic may be limited, if the power law relation between initial learning rate and model error for SGD ANNs with a rectified linear activation function is reliable. This finding indicates that smaller changes in synaptic weights in the neural network facilitate better performance, which when analysed within the context of maximum iteration limit of 200, may indicate that these relationships do not significantly change over time.

# Conclusion

The aim of this paper is to contribute to the understanding of the nature of optimal road traffic forecasting models for the general case to better inform practitioners. This was in response to a lack of statistically robust comparisons of model types in the literature.

It is envisaged that such a study has not been completed before due to the recent advent of accessible cloud computing such as AWS and the development of open-source machine learning libraries such as scikit-learn (Pedregosa, et al., 2011).

Through the development of a randomised hyperparameter sweep conducted on a powerful 32 core machine, 186880 traffic forecasting models were developed and tested. K-means clustering of these models and reporting of their median performance facilitated an analysis of the role of quantitative and qualitative hyperparameters for the generalised for the roads within the study sample, which consisted of 20 randomly selected roads in Aarhus, Denmark in 2014.

Multi-layer perceptron ANNs were found to have lower error, on average, that SVMs. For ANNs, the L-BFGS learning algorithm showed the best performance with all activation functions performing as well as one another for this model species, except for the hyperbolic tangent function which outperformed the rectified linear function. For SVMs, the linear kernel outperformed all others.

Through the development of a binned minimum plot, trends in hyperparameters were able to be understood in the case that other hyperparameters were optimised. This led to insights into the relationships between model complexity and error, the need for sufficient learning iterations, and into the relationships between model error and both momentum and learning rates in the case of ANNs and model error and regression boundary widths ($\varepsilon$) in the case of SVMs.

If the data is representative of traffic generally, the findings of these report pertain to the general case.

The sample size of this study was limited due to cost constraints, and the extent of hyperparameters examined was limited due to human error. With more data, resources and learnings from this paper, practitioners will be able to extend the scope of the analysed hyperparameter feature space and extend the applicability of such findings.

This study represents what the advent of cloud computing and open source technology can offer the field of transportation analytics, time series forecasting and machine learning generally. Should research in this and related fields follow this trajectory, both machine learning practitioners and society stand to benefit from having reliable findings pertaining to the nature of supervised machine learning models and their hyperparameters.

# Bibliography


Adeli, H., 2001. Neural networks in civil engineering: 1989-2000. *Computer-Aided Civil and Infrastructure Engineering,* 16(2), pp. 126-142.

Armstrong, J. S., 2001. Evaluating Forecasting Methods. In: *Principles of Forecasting: A Handbook for Researchers and Practitioners.* s.l.:Kluwer Academic Publishers.

Baydin, A. G. et al., 2017. *Online Learning Rate Adaptation with Hypergradient Descention.,* s.l.: s.n.

Bergstra, J. & Bengio, Y., 2012. Random Search for Hyperparameter Optimisation. *Journal of Machine Learning Research,* Volume 13, pp. 281-305.

Bloksgaard, M. & Christiansen, A. K., 2015. *Use of travel time data from citywide Bluetooth system.* Bordeaux, The Transportation Research Board.

Boser, B. E., Guyon, I. M. & Vapnik, V. N., 1992. *A Training Algorithm for Optimal Marginal Classifiers.* s.l., s.n., pp. 144-152.

Buscema, M., 1998. Back Propagation Neural Networks. *Substance Use & Misuse,* 33(2), pp. 233-270.

Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K. & Han, L. D., 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert Systems with Applications,* Volume 36, p. 6164–6173.

Chandra, S. R. & Al-Deek, H., 2009. Predictions of Freeway Traffic Speeds and Volumes Using Vector Autoregressive Models. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations,* 13(2), pp. 53-72.

Ding, A., Zhao, X. & Jiao, L., 2002. *TRAFFIC FLOW TIME SERIES PREDICTIONBASED ON STATISTICS LEARNING THEORY.* Singapore, IEEE.

Ghasemi, A. & Zahediasl, S., 2012. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism,* 10(2), pp. 486-489.

Habtie, A. B., Abraham, A. & Midekso, D., 2017. Artificial Neural Network Based Real-Time Urban Road Traffic State Estimation Framework. In: A. Abraham, R. Falcon & M. Koeppen, eds. *Computational Intelligence in Wireless Sensor Networks* . s.l.:Springer, pp. 73-97.

Haworth, J., 2014. *Spatiotemporal forecasting of network data.* London: University College London.

Karlaftis, M. & Vlahogianni, E., 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C.*

Kingma, D. P. & Ba, J. L., 2015. *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION.* San Diego, s.n.

LeCun, Y., Bottou, L., Orr, G. B. & Müller, K.-R., 1998. *Efficient BackProp,* Red Bank, NJ: Springer.

Li, R. & Rose, G., 2011. Incorporating uncertainty into short-term travel time predictions. *Transportation Research Part C,* 19(6), pp. 1006-1018.

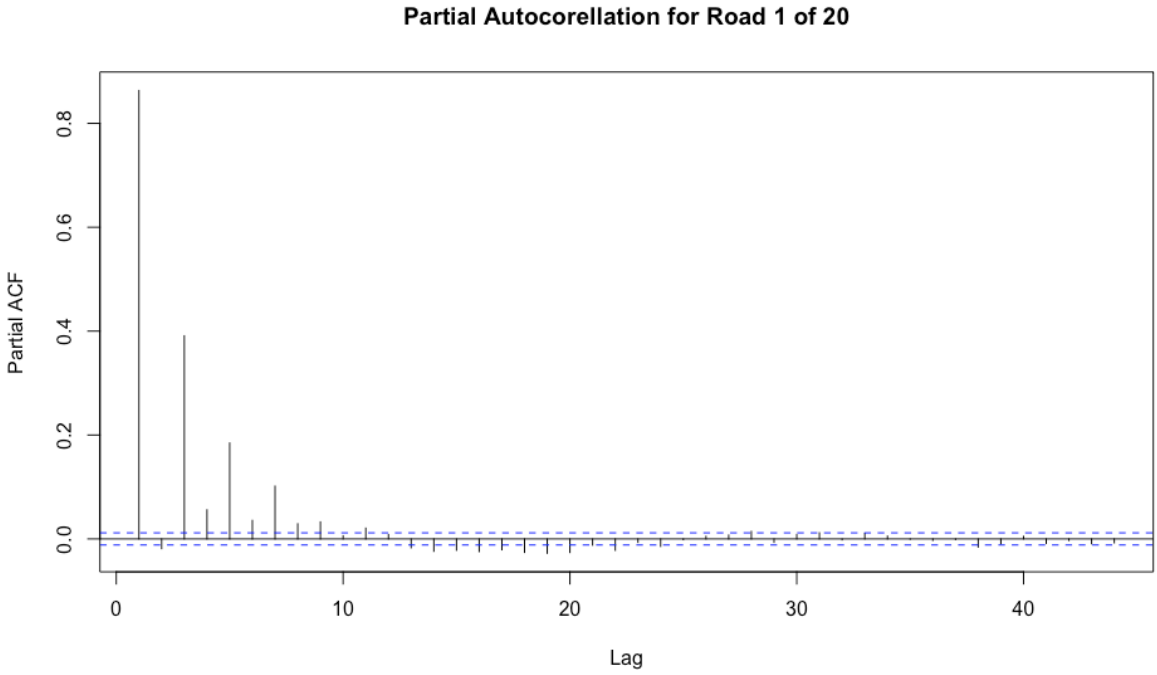
Liu, X. et al., 2011. A Short-Term Forecasting Algorithm for Network Traffic Based on Chaos Theory and SVM. *Journal of Network System Management,* Volume 19, pp. 427-447.

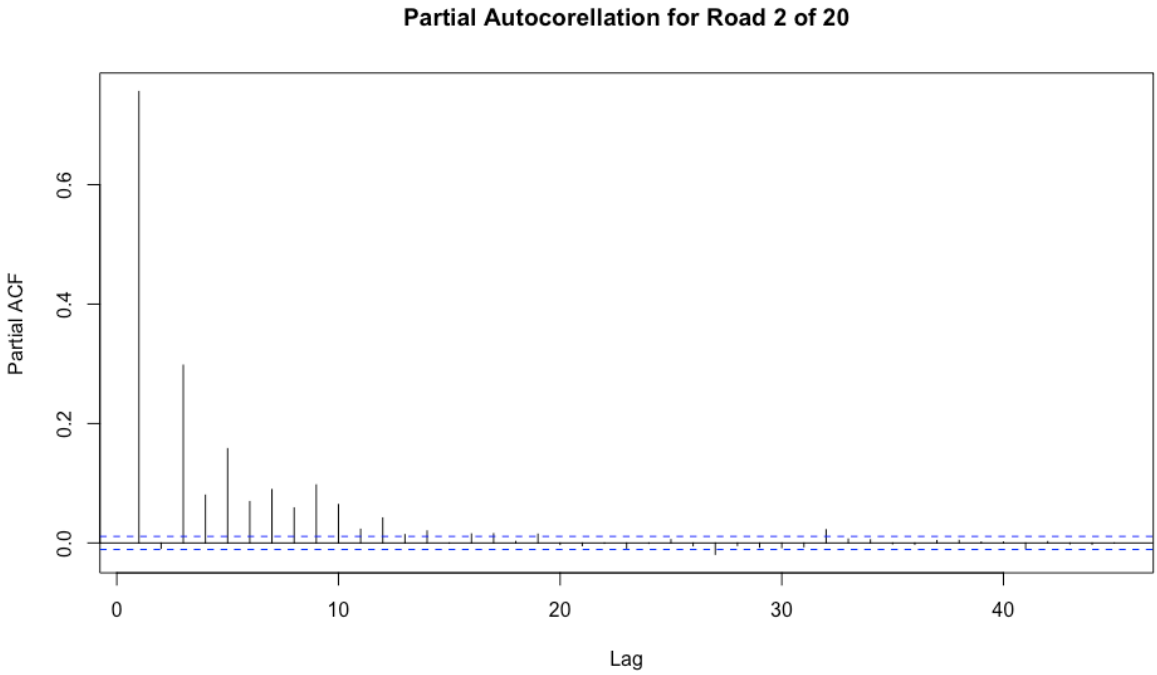
Lv, Y. et al., 2015. Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS,* April, 16(2), pp. 865-873.

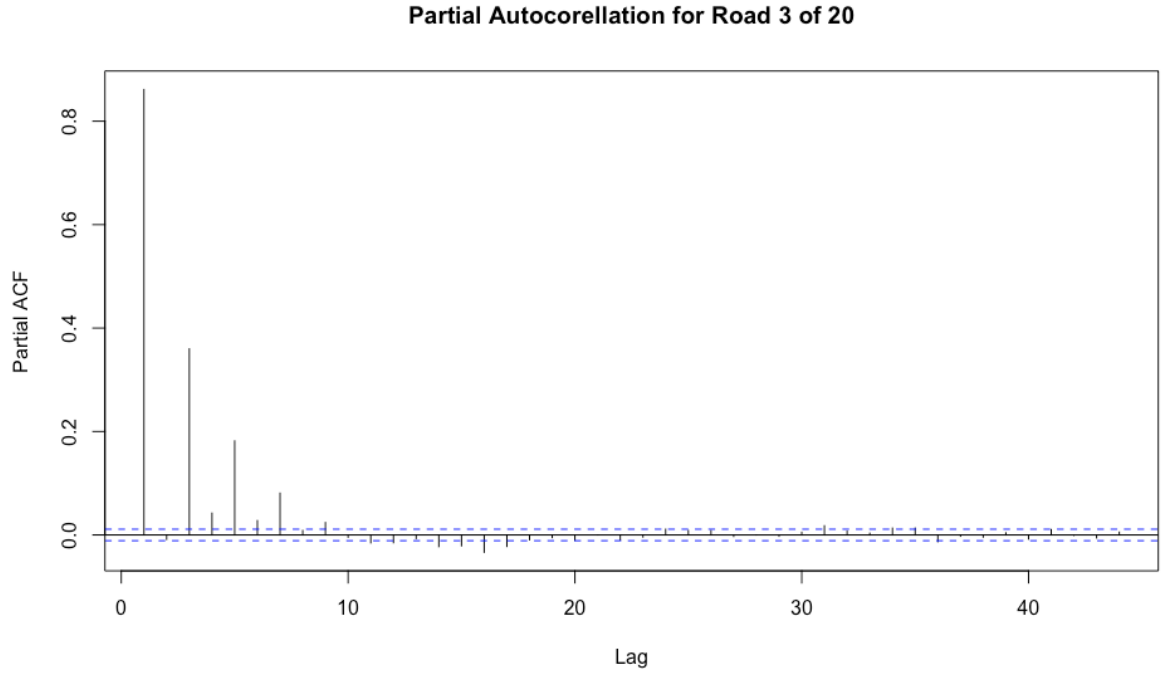
Min, W. & Wynter, L., 2011. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies,* 19(4), pp. 606-616.
Moretti, F., Pizzuti, S., Panzieri, S. & Annunziato, M., 2015. Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling. *Neurocomputing,* pp. 3-7.
Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* Volume 12, pp. 2825--2830.
Rumelhart, D. E., Hinton, G. E. & Williams, R. J., 1986. Learning Representations by Backpropagation Errors. *Nature,* 9 October, Volume 323, pp. 533-536.
Smola, A. J. & Schölkopf, B., 2004. A tutorial in support vector regression. *Statistics and Computation,* Volume 14, pp. 199-222.
Stathopoulos, A., Dimitriou, L. & Tsekeris, T., 2008. Fuzzy modeling approach for combined forecasting of urban traffic flow. *Computer-aided Civil and Infrastructure Engineering ,* Volume 23, pp. 521-535.
Vlahogianni, E. I., Golias, J. C. & Karlaftis, M. G., 2004. Short-term traffic forecasting: Overview of objectives and methods. *Transport Reviews,* 24(5), pp. 533-557.
Vlahogianni, E. I., Karlaftis, M. G. & Golias, J. C., 2014. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C.*
Wickham, H., 2009. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag.
Williams, M. N., Grajales, C. A. G. & Krukiewicz, D., 2013. Assumptions for Multiple Regression: Correcting Two Misconceptions. *Practical Assessment, Research & Evaluation,* 18(11), pp. 1-14.
Wu, C.-H., Ho, J.-M. & Lee, D. T., 2004. Travel-Time Prediction With Support Vector Regression. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS,* 5(4), pp. 276-281.
Yao, X. & Lin, Y., 1998. Towards designing artificial neural networks by evolution. *Applied Mathematics and Computation,* Volume 91, pp. 83-90.
Yasdi, R., 1999. Prediction of Road Traffic Using a Neural Network Approach. *Neural Computing and Applications,* Volume 8, pp. 135-142.
Zheng, W., Lee, D.-H. & Shi, Q., 2006. Short-Term Freeway Traffic Flow Prediction: Bayesian Combined Neural Network Approach. *Journal of Transport Engineering,* 132(2), pp. 114-121.

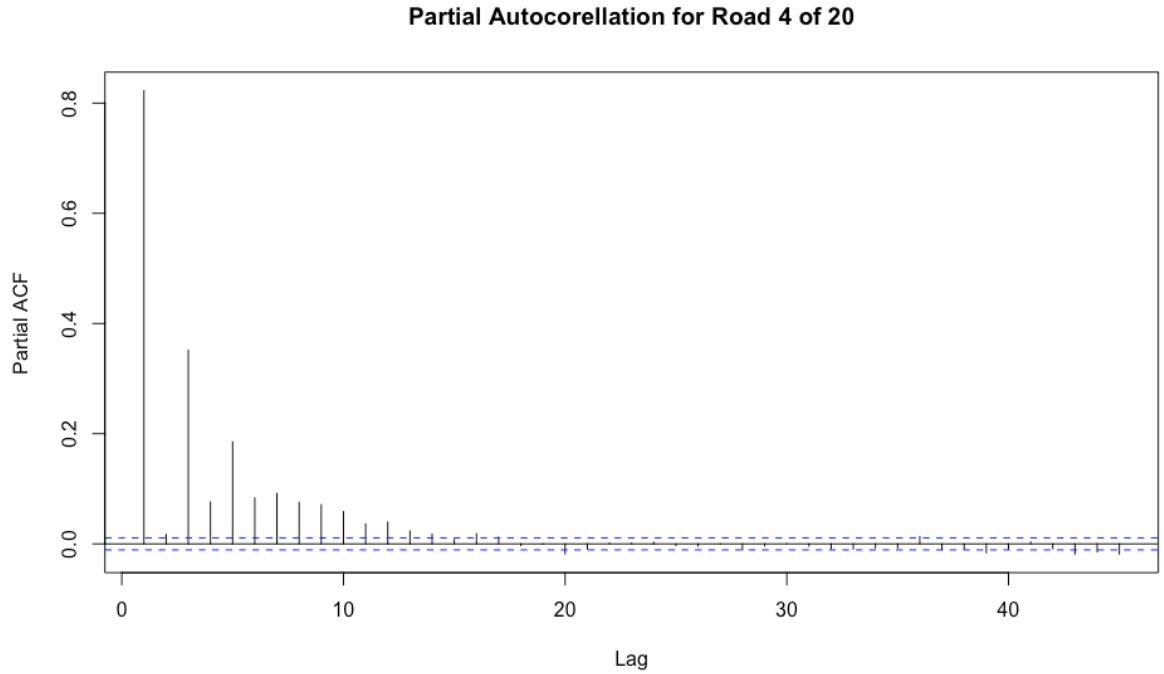
# Appendix A: PACF Plots for Random Road Segments.

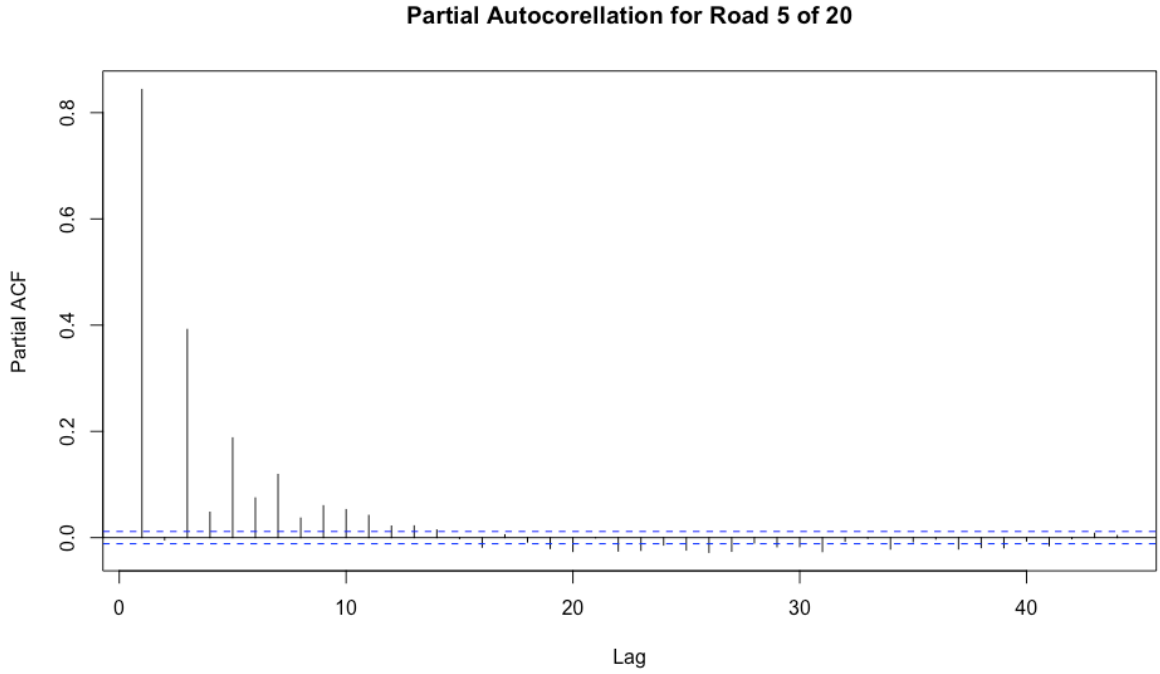
All plots in this appendix were created with R and ggplot2 (Wickham, 2009).

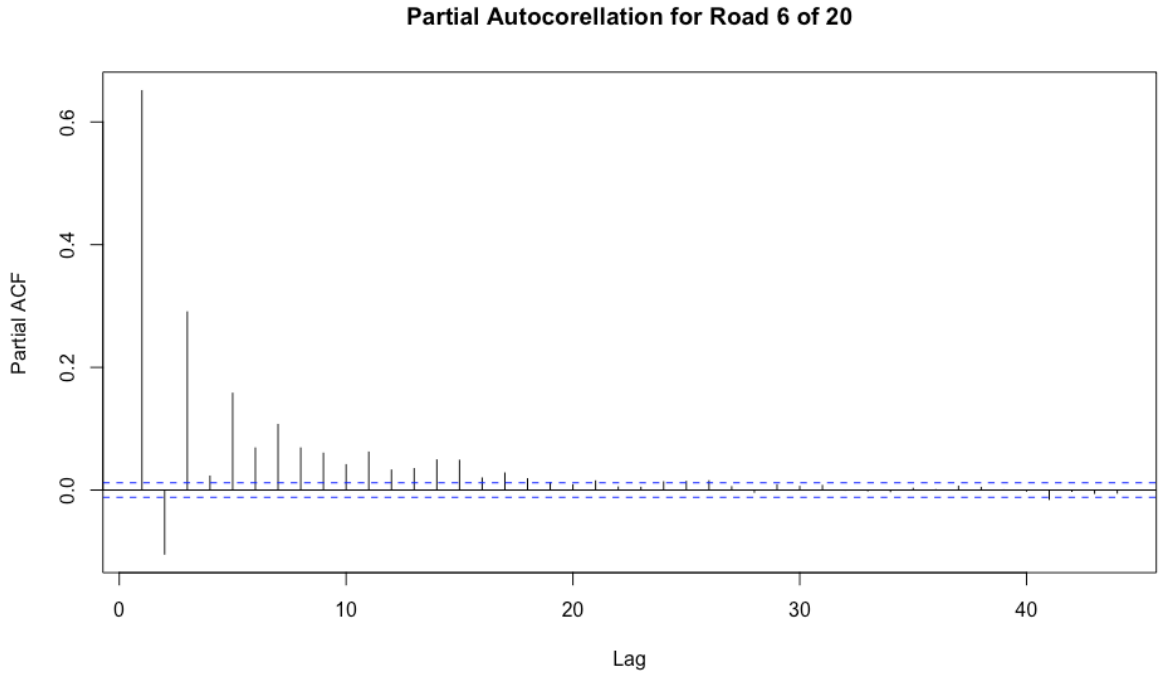


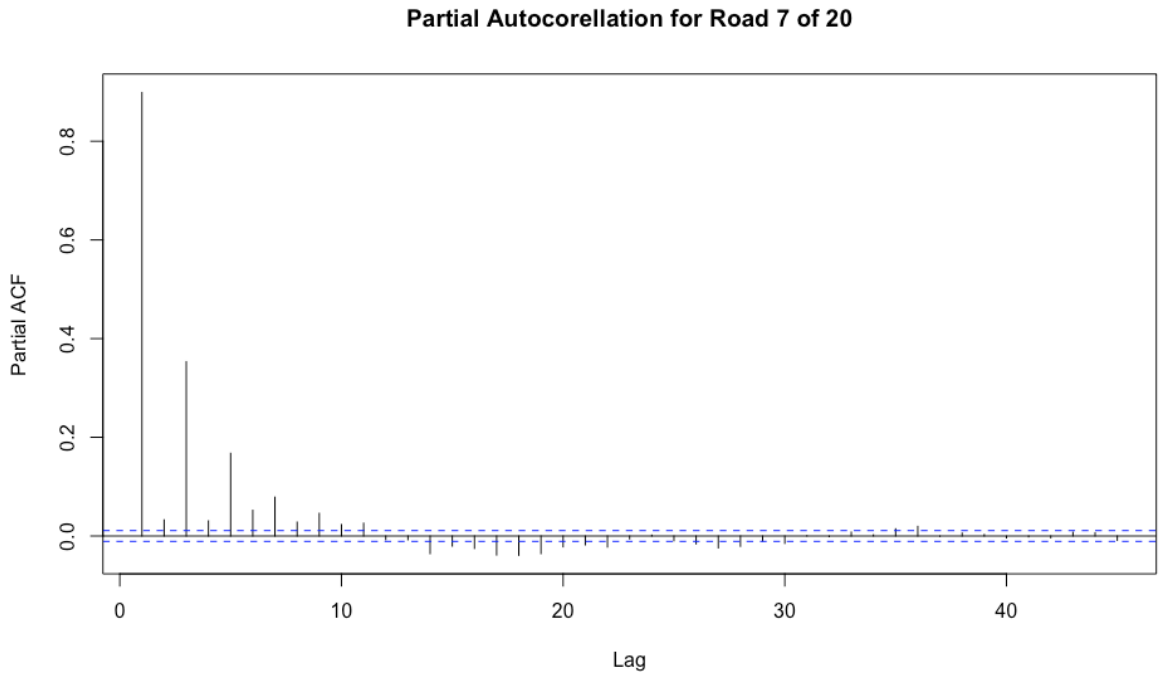



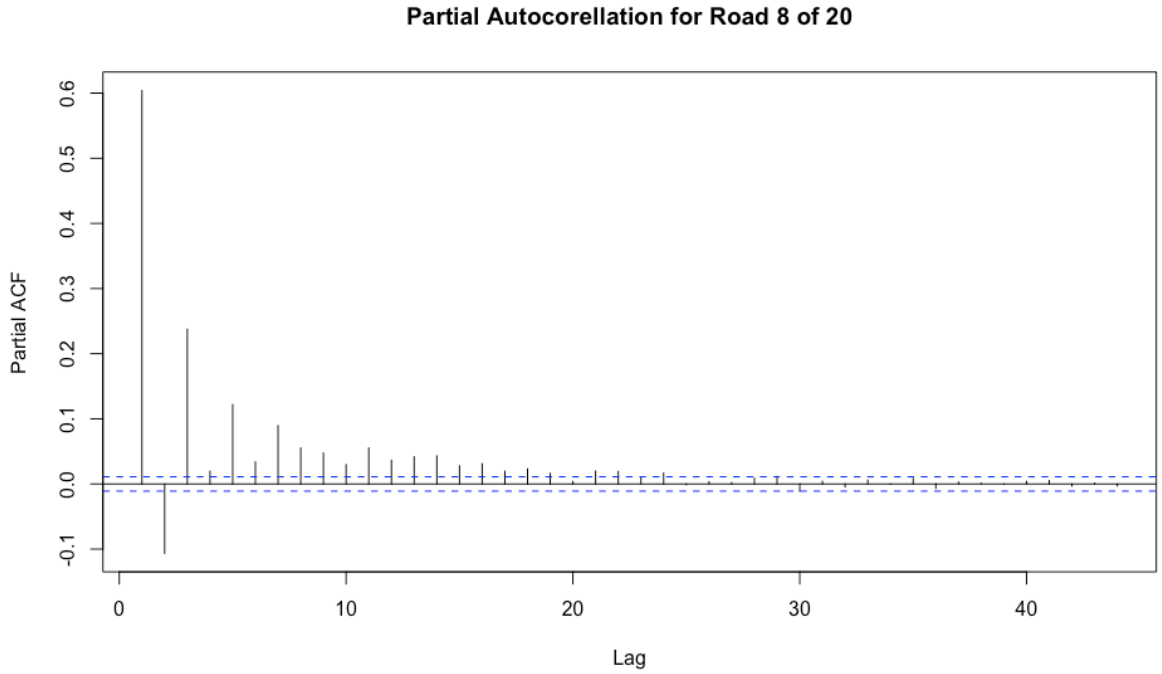

Partial Autocorellation for Road 3 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag

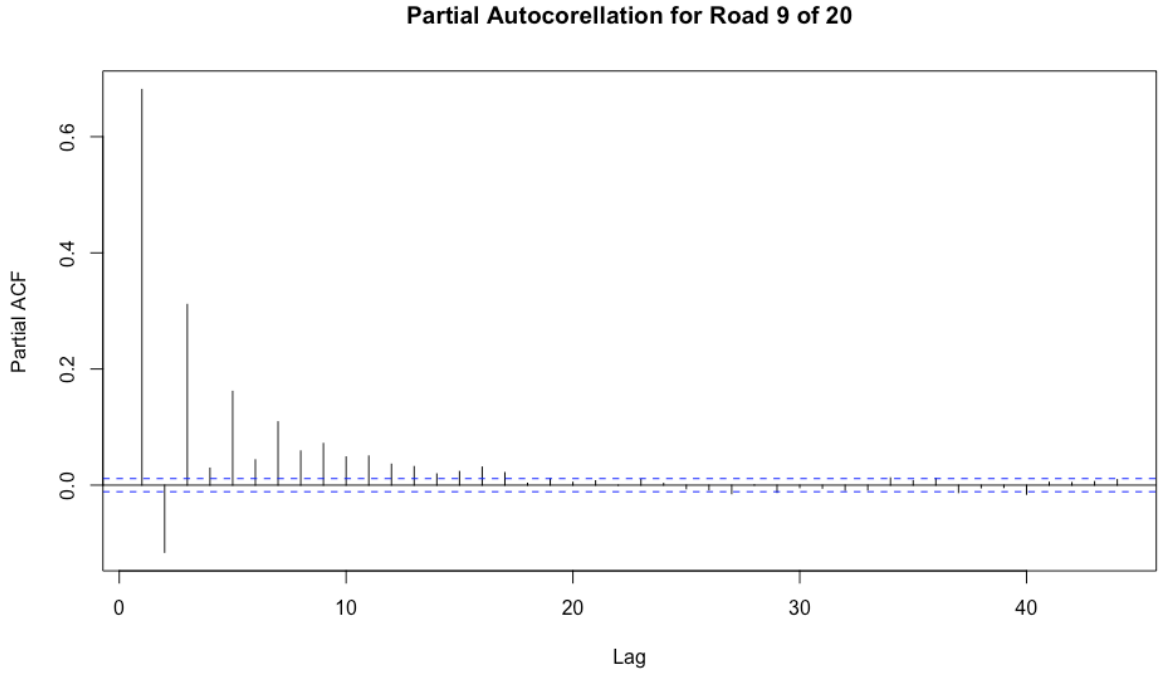

Partial Autocorellation for Road 4 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag

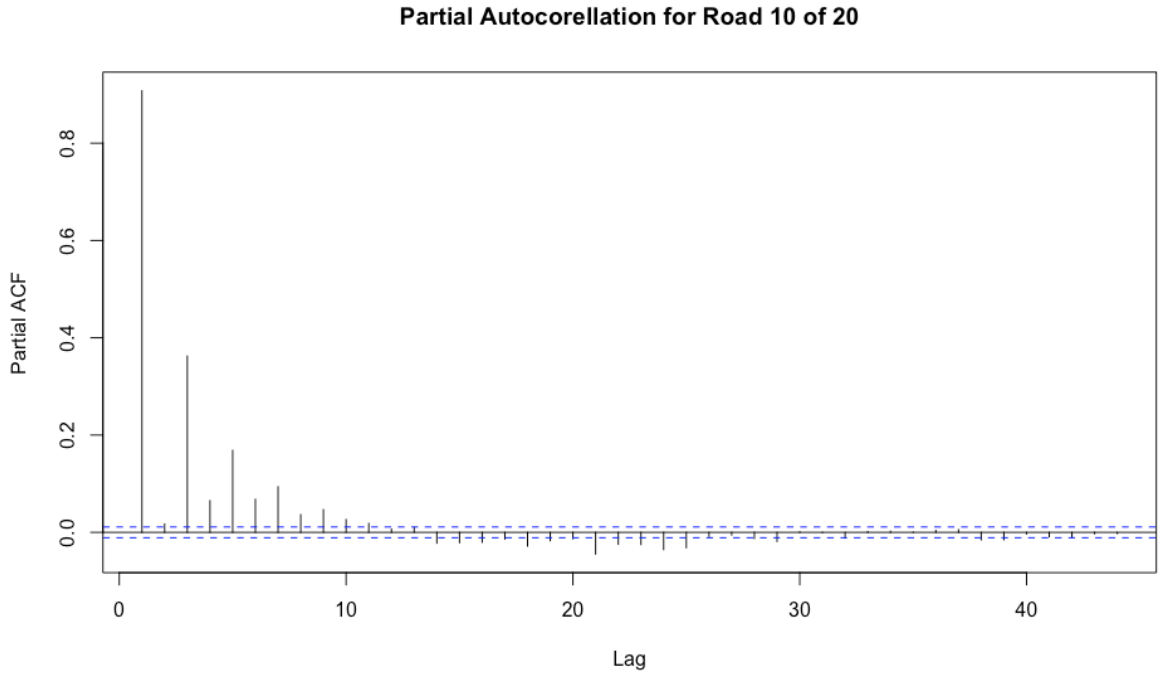

Partial Autocorellation for Road 5 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag

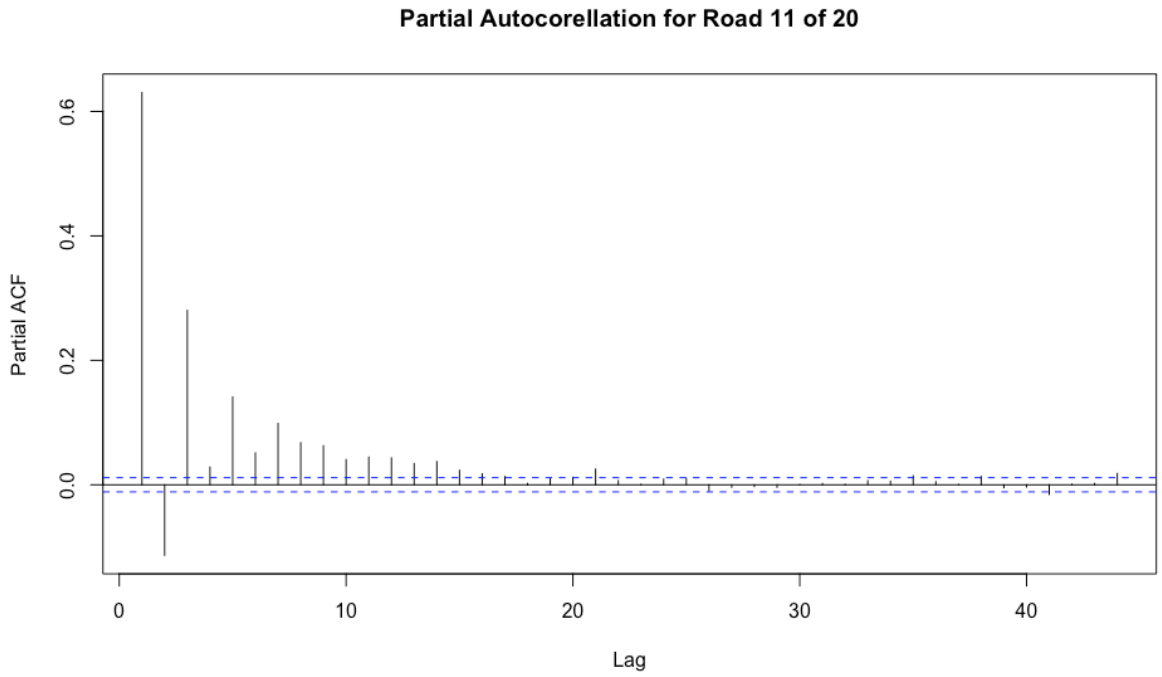

Partial Autocorellation for Road 6 of 20
Partial ACF
0.0
0.2
0.4
0.6
0
10
20
30
40
Lag

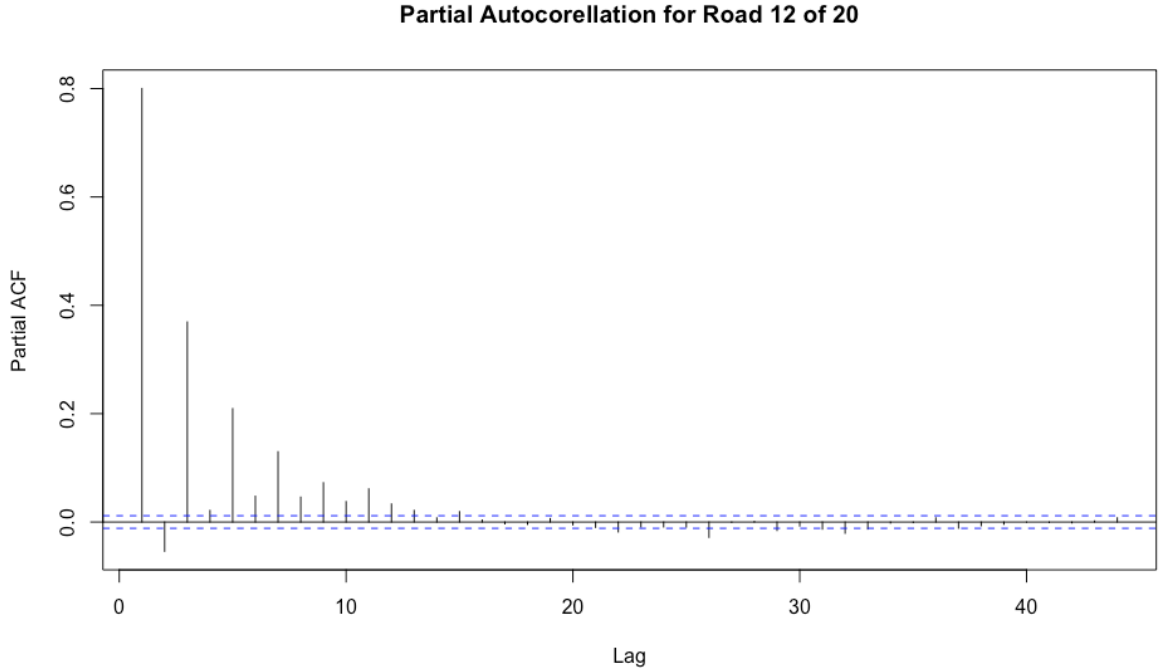

Partial Autocorellation for Road 7 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag

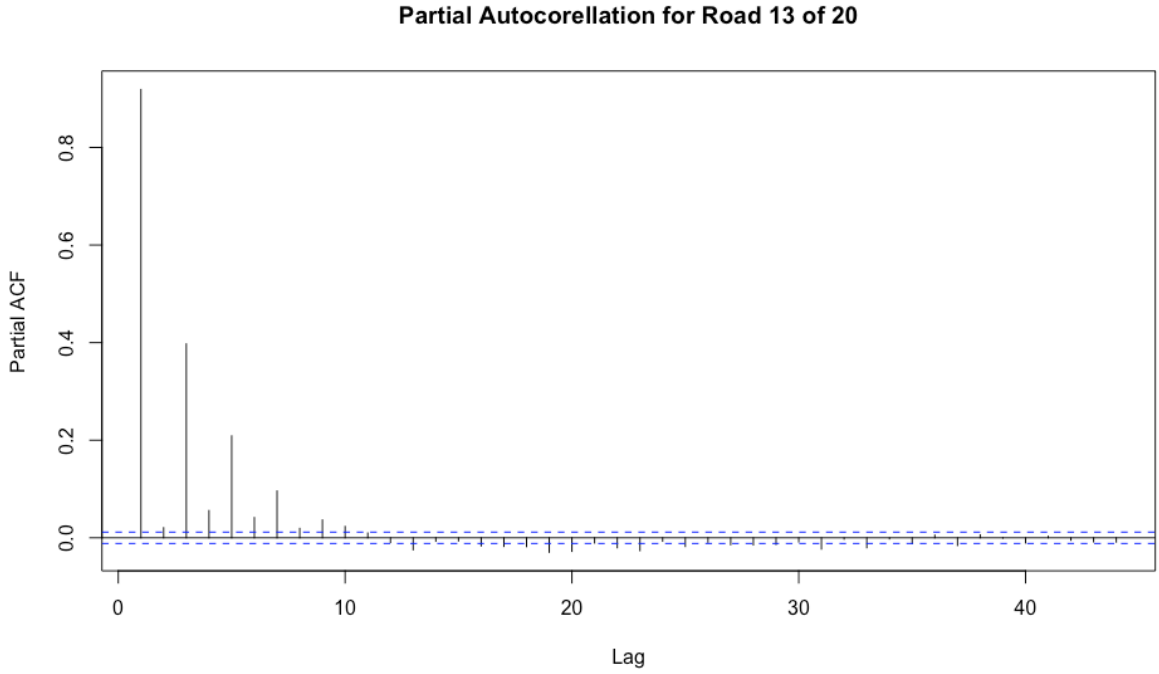

Partial Autocorellation for Road 8 of 20
Partial ACF
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
10
20
30
40
Lag

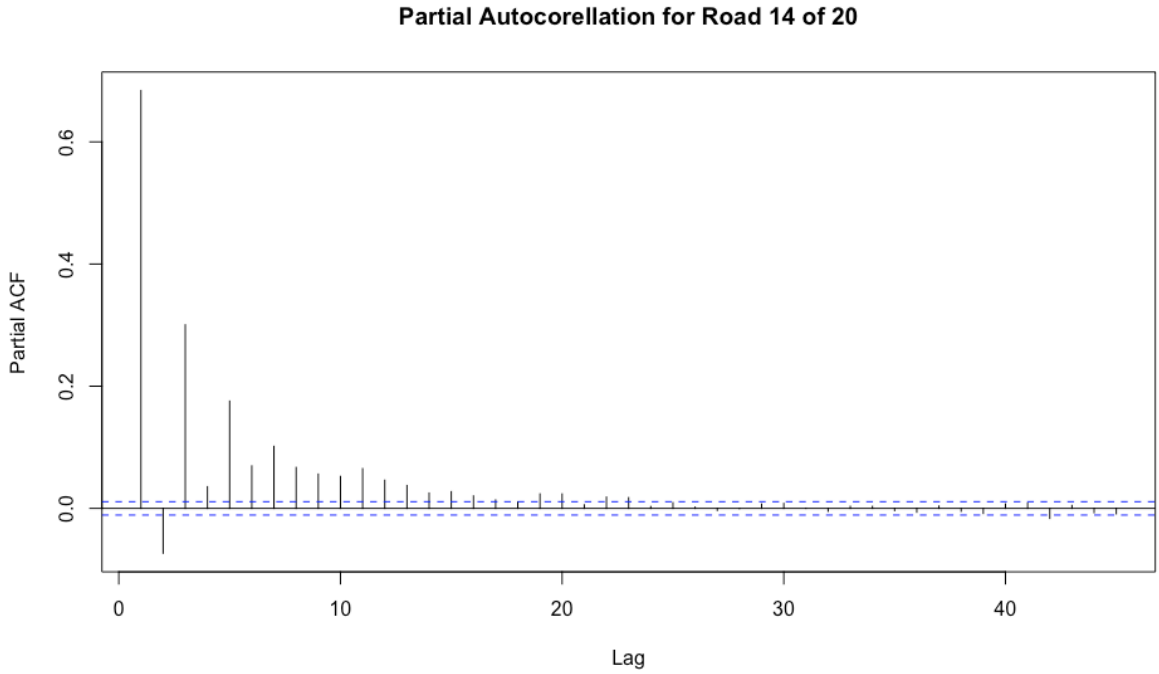

Partial Autocorellation for Road 9 of 20
Partial ACF
0.0
0.2
0.4
0.6
0
10
20
30
40
Lag

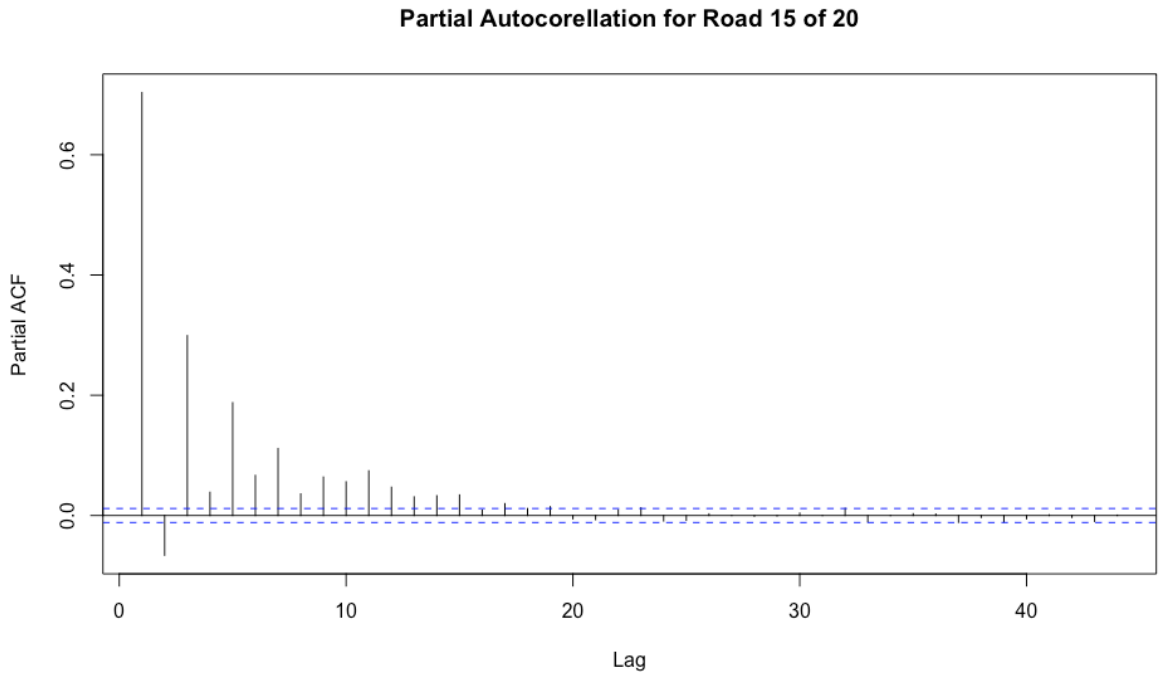

Partial Autocorellation for Road 10 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag

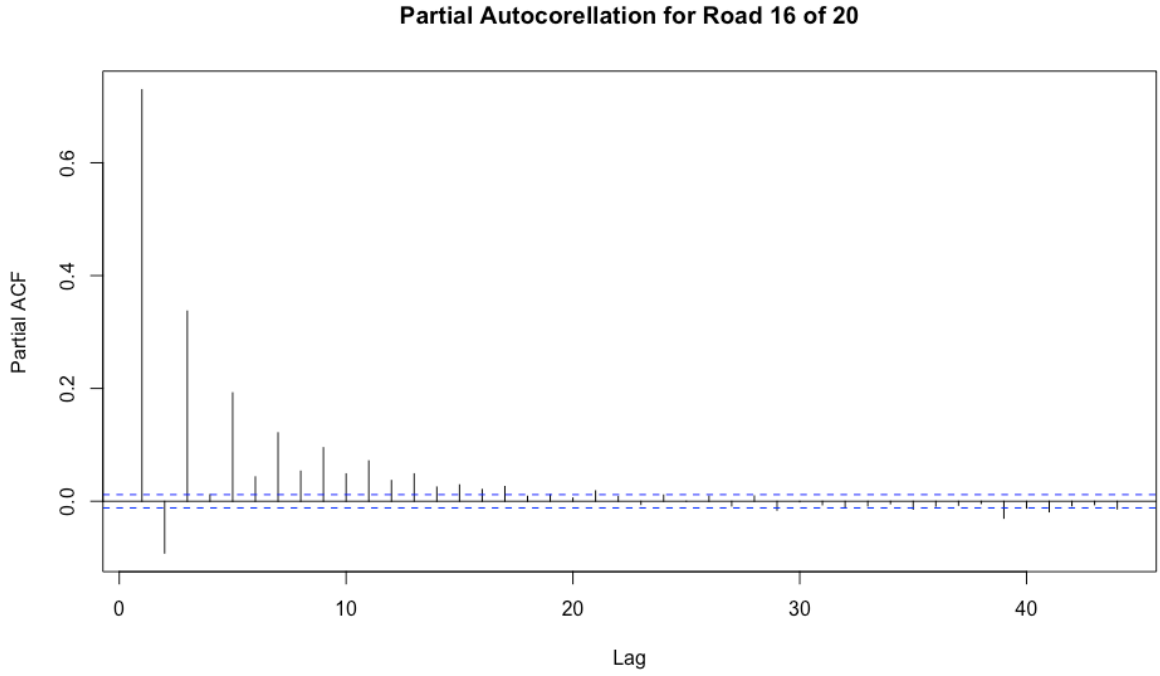

Partial Autocorellation for Road 11 of 20
Partial ACF
0.0
0.2
0.4
0.6
0
10
20
30
40
Lag

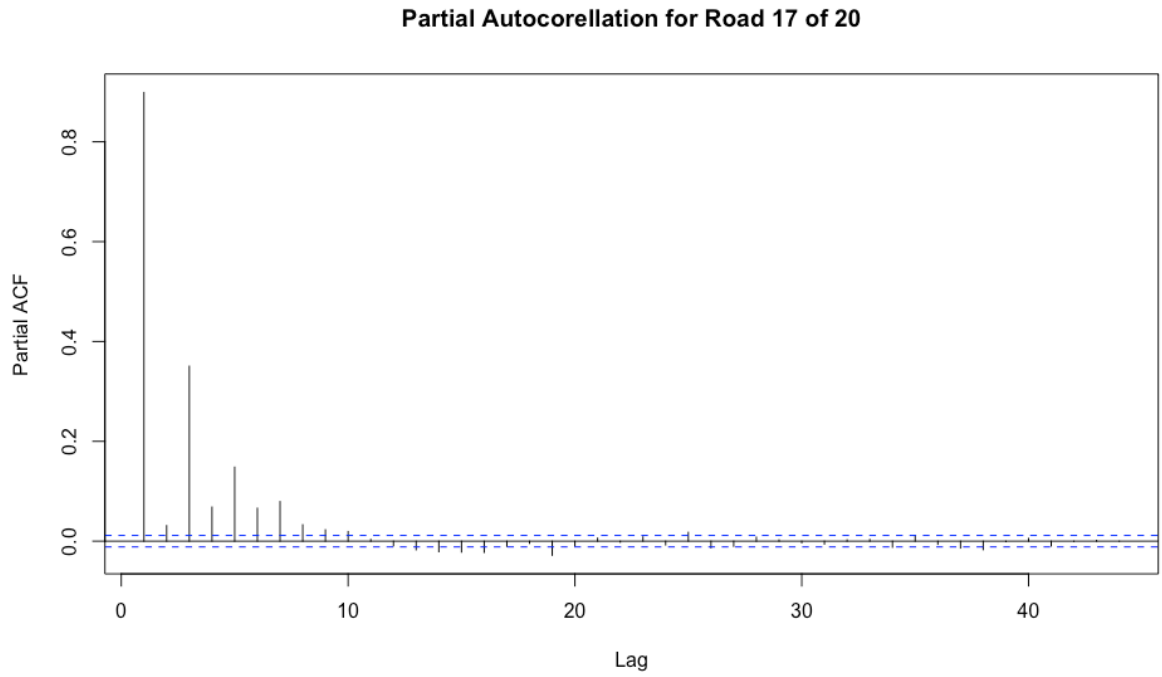

Partial Autocorellation for Road 12 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag

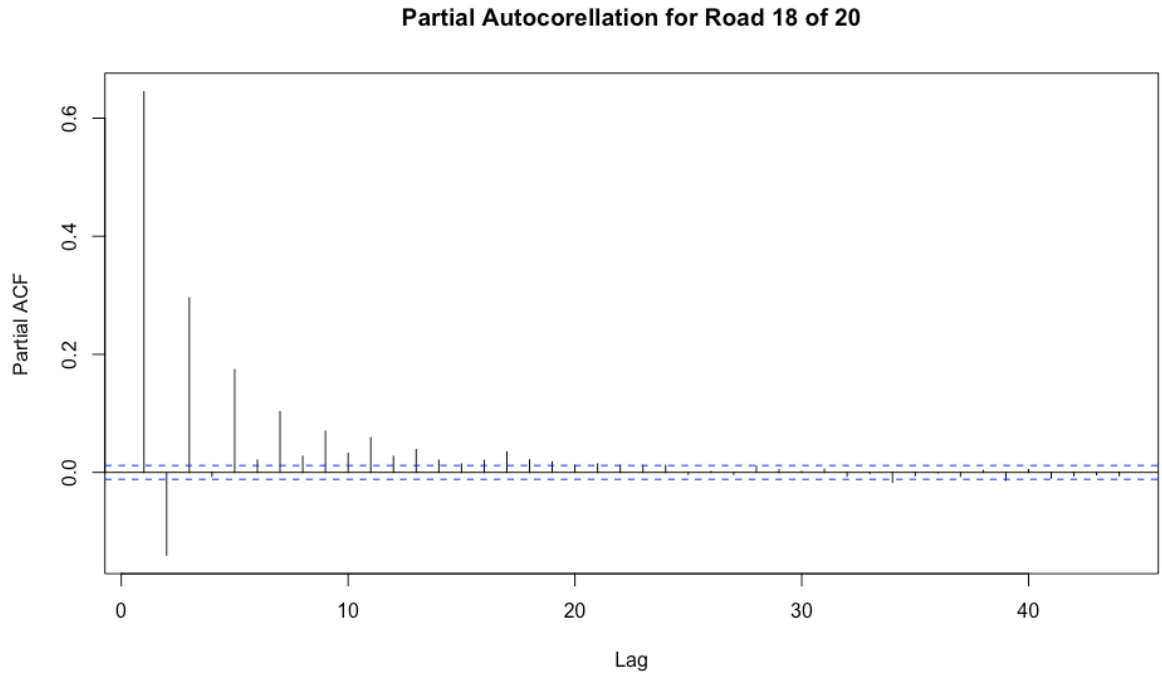

Partial Autocorellation for Road 13 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag

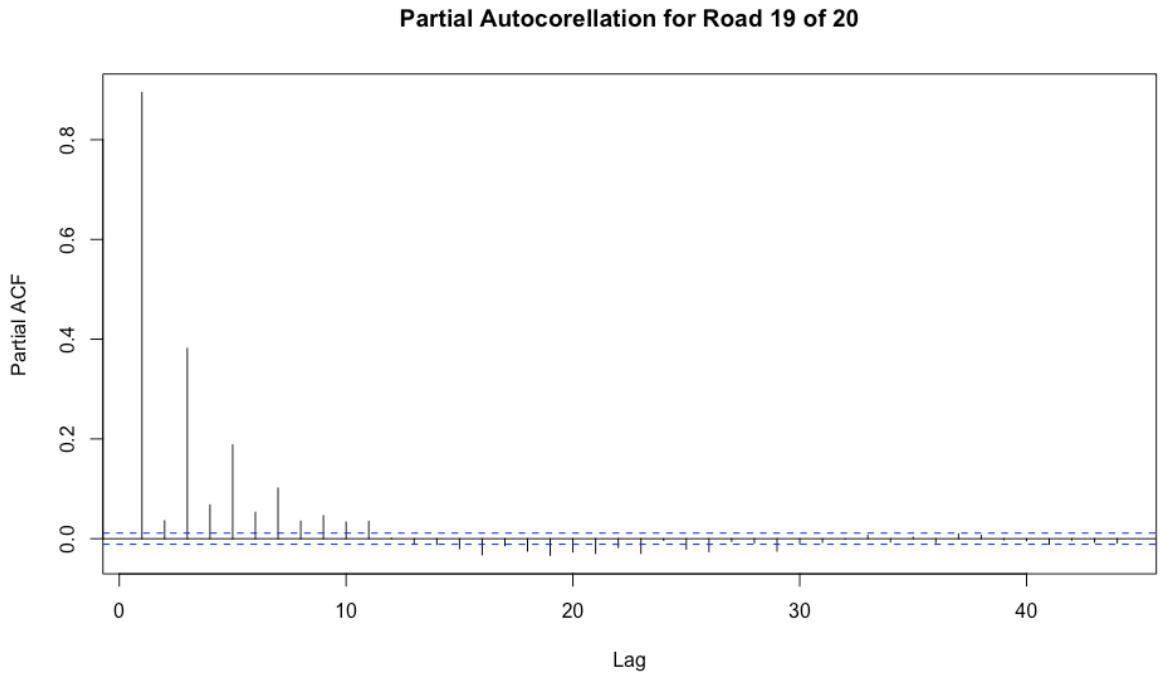

Partial Autocorellation for Road 14 of 20
Partial ACF
0.0
0.2
0.4
0.6
0
10
20
30
40
Lag

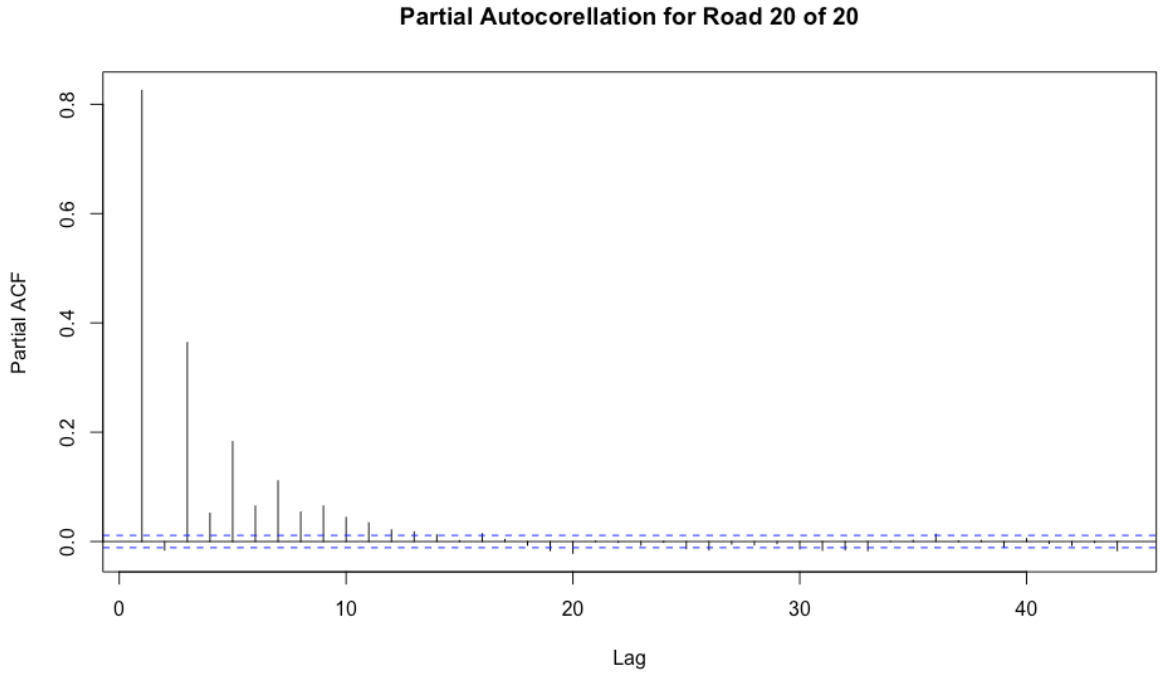

Partial Autocorellation for Road 15 of 20
Partial ACF
0.0
0.2
0.4
0.6
0
10
20
30
40
Lag


Partial Autocorellation for Road 16 of 20
Partial ACF
0.0
0.2
0.4
0.6
0
10
20
30
40
Lag


Partial Autocorellation for Road 17 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag


Partial Autocorellation for Road 18 of 20
Partial ACF
0.0
0.2
0.4
0.6
0
10
20
30
40
Lag


Partial Autocorellation for Road 19 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag


Partial Autocorellation for Road 20 of 20
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
Lag

# Appendix B: Welch's t-test Results

## B.1 Genera Analysis

*SVM vs ANN*

| **Error Metric** | $\overline{x}_{ANN}$ | $\overline{x}_{SVM}$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|
| MAE | 5.26017 | 10.52113 | 0.001318099 |
| MEDAE | 4.426786 | 9.65361 | 0.0004454453 |

*Spatial Lag Analysis*

| **Error Metric** | **Spatial Lag i** | **Spatial Lag j** | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 10.22412 | 8.421395 | 0.006083975 |
| MEDAE | 0 | 1 | 11.8131 | 7.730697 | 7.348043e-07 |

*Temporal Lag Analysis All Models*

| **Error Metric** | **Temporal Lag i** | **Temporal Lag j** | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 9.745047 | 9.10548 | 0.489526 |
| MAE | 2 | 4 | 9.745047 | 13.81907 | 0.003274115 |
| MAE | 2 | 5 | 9.745047 | 10.7357 | 0.3859126 |
| MAE | 3 | 4 | 9.10548 | 13.81907 | 0.0008611607 |
| MAE | 3 | 5 | 9.10548 | 10.7357 | 0.1662062 |
| MAE | 4 | 5 | 13.81907 | 10.7357 | 0.04872985 |
| MEDAE | 2 | 3 | 8.391366 | 9.46359 | 0.1544502 |
| MEDAE | 2 | 4 | 8.391366 | 10.93041 | 0.006799354 |
| MEDAE | 2 | 5 | 8.391366 | 11.1803 | 0.005683319 |
| MEDAE | 3 | 4 | 9.46359 | 10.93041 | 0.1900545 |
| MEDAE | 3 | 5 | 9.46359 | 11.1803 | 0.1453318 |
| MEDAE | 4 | 5 | 10.93041 | 11.1803 | 0.8480745 |

*Temporal ANN*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 6.736882543 | 5.574116105 | 0.306814573 |
| MAE | 2 | 4 | 6.736882543 | 4.729921101 | 0.067234544 |
| MAE | 2 | 5 | 6.736882543 | 3.471162532 | 0.000484636 |
| MAE | 3 | 4 | 5.574116105 | 4.729921101 | 0.449084245 |
| MAE | 3 | 5 | 5.574116105 | 3.471162532 | 0.028092125 |
| MAE | 4 | 5 | 4.729921101 | 3.471162532 | 0.165628762 |
| MEDAE | 2 | 3 | 5.870360458 | 4.533515661 | 0.193475159 |
| MEDAE | 2 | 4 | 5.870360458 | 4.235688781 | 0.122446843 |
| MEDAE | 2 | 5 | 5.870360458 | 2.577484248 | 7.56E-05 |
| MEDAE | 3 | 4 | 4.533515661 | 4.235688781 | 0.772894817 |
| MEDAE | 3 | 5 | 4.533515661 | 2.577484248 | 0.014233015 |
| MEDAE | 4 | 5 | 4.235688781 | 2.577484248 | 0.047434138 |

*Spatial ANN*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 3.88189 | 6.000191 | 0.004380585 |
| MEDAE | 0 | 1 | 3.442839 | 4.955087 | 0.03202771 |

*Temporal SVM*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 8.428554945 | 7.076770665 | 0.482974219 |
| MAE | 2 | 4 | 8.428554945 | 13.51904808 | 0.295067627 |
| MAE | 2 | 5 | 8.428554945 | 11.245904 | 0.521549518 |
| MAE | 3 | 4 | 7.076770665 | 13.51904808 | 0.156586235 |
| MAE | 3 | 5 | 7.076770665 | 11.245904 | 0.302427852 |
| MAE | 4 | 5 | 13.51904808 | 11.245904 | 0.705185289 |
| MEDAE | 2 | 3 | 7.104604396 | 7.460311927 | 0.604554859 |
| MEDAE | 2 | 4 | 7.104604396 | 10.86719231 | 0.245399792 |
| MEDAE | 2 | 5 | 7.104604396 | 11.612004 | 0.261830172 |
| MEDAE | 3 | 4 | 7.460311927 | 10.86719231 | 0.299090915 |
| MEDAE | 3 | 5 | 7.460311927 | 11.612004 | 0.305333566 |
| MEDAE | 4 | 5 | 10.86719231 | 11.612004 | 0.884463186 |

*Spatial SVM*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 10.5484961 | 7.529273077 | 0.126575828 |
| MEDAE | 0 | 1 | 8.95165625 | 8.190151923 | 0.597417241 |

## B.2 Species Analysis: ANN

*Learning Functions*

| Error Metric | Learning Function i | Learning Function j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | lbfgs | adam | 1.215478702 | 10.36071685 | 2.75E-25 |
| MAE | lbfgs | sgd | 1.215478702 | 3.308344986 | 0.001049337 |
| MAE | sgd | adam | 3.308344986 | 10.36071685 | 7.56E-11 |
| MEDAE | lbfgs | adam | 0.776814924 | 8.87795635 | 1.02E-24 |
| MEDAE | lbfgs | sgd | 0.776814924 | 2.920791466 | 0.000791484 |
| MEDAE | sgd | adam | 2.920791466 | 8.87795635 | 3.91E-09 |

### *B.2.1 L-BFGS*

*Activation Functions*

| Error Metric | Activation Function i | Activation Function j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | identity | logistic | 1.17041719 | 1.164663755 | 0.811108642 |
| MAE | identity | tanh | 1.17041719 | 1.223606504 | 0.041940205 |
| MAE | identity | relu | 1.17041719 | 1.320937198 | 1.67E-08 |
| MAE | logistic | tanh | 1.164663755 | 1.223606504 | 0.033057788 |
| MAE | logistic | relu | 1.164663755 | 1.320937198 | 2.89E-08 |
| MAE | tanh | relu | 1.223606504 | 1.320937198 | 0.00111706 |
| MEDAE | identity | logistic | 0.72674843 | 0.735939592 | 0.60206039 |
| MEDAE | identity | tanh | 0.72674843 | 0.761474797 | 0.074171735 |
| MEDAE | identity | relu | 0.72674843 | 0.899158213 | 1.05E-15 |
| MEDAE | logistic | tanh | 0.735939592 | 0.761474797 | 0.213629106 |
| MEDAE | logistic | relu | 0.735939592 | 0.899158213 | 3.73E-13 |
| MEDAE | tanh | relu | 0.761474797 | 0.899158213 | 7.90E-09 |

*Spatial Lags*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 1.229678005 | 1.200205793 | 0.125389361 |
| MEDAE | 0 | 1 | 0.788077098 | 0.76470122 | 0.114214562 |

*Temporal Lags*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 1.260021324 | 1.243932984 | 0.556649967 |
| MAE | 2 | 4 | 1.260021324 | 1.154798299 | 0.000151389 |
| MAE | 2 | 5 | 1.260021324 | 1.198457933 | 0.025233083 |
| MAE | 3 | 4 | 1.243932984 | 1.154798299 | 0.000633037 |
| MAE | 3 | 5 | 1.243932984 | 1.198457933 | 0.078125381 |
| MAE | 4 | 5 | 1.154798299 | 1.198457933 | 0.0953747 |
| MEDAE | 2 | 3 | 0.80459596 | 0.792585664 | 0.579688408 |
| MEDAE | 2 | 4 | 0.80459596 | 0.738566829 | 0.002388613 |
| MEDAE | 2 | 5 | 0.80459596 | 0.768634615 | 0.090575084 |
| MEDAE | 3 | 4 | 0.792585664 | 0.738566829 | 0.007343312 |
| MEDAE | 3 | 5 | 0.792585664 | 0.768634615 | 0.222260432 |
| MEDAE | 4 | 5 | 0.738566829 | 0.768634615 | 0.126081344 |

*B.2.2 SGD*

*Activation Functions*

| Error Metric | Activation Function i | Activation Function j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | identity | logistic | 1.223282955 | 5.989610837 | 0.0094773 |
| MAE | identity | tanh | 1.223282955 | 2.649614183 | 0.237325493 |
| MAE | identity | relu | 1.223282955 | 2.583498358 | 0.097914902 |
| MAE | logistic | tanh | 5.989610837 | 2.649614183 | 0.127921136 |
| MAE | logistic | relu | 5.989610837 | 2.583498358 | 0.089964561 |
| MAE | tanh | relu | 2.649614183 | 2.583498358 | 0.963834646 |
| MEDAE | identity | logistic | 0.785097727 | 5.691650246 | 0.007589315 |
| MEDAE | identity | tanh | 0.785097727 | 2.20525601 | 0.239475928 |
| MEDAE | identity | relu | 0.785097727 | 2.181735632 | 0.089304242 |
| MEDAE | logistic | tanh | 5.691650246 | 2.20525601 | 0.112149497 |
| MEDAE | logistic | relu | 5.691650246 | 2.181735632 | 0.080675012 |
| MEDAE | tanh | relu | 2.20525601 | 2.181735632 | 0.987134932 |

*Rate Functions*

| Error Metric | Rate Function i | Rate Function j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | constant | adaptive | 2.983063 | 3.693953 | 0.5859653 |
| MEDAE | constant | adaptive | 2.576893 | 3.328468 | 0.564851 |

*Spatial Lags*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 2.635933794 | 3.525334821 | 0.481331757 |
| MEDAE | 0 | 1 | 2.264500988 | 3.132579082 | 0.492045342 |

*Temporal Lags*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 3.719530714 | 2.681030146 | 0.520711313 |
| MAE | 2 | 4 | 3.719530714 | 4.045828502 | 0.876983442 |
| MAE | 2 | 5 | 3.719530714 | 2.699973904 | 0.529649457 |
| MAE | 3 | 4 | 2.681030146 | 4.045828502 | 0.494485562 |
| MAE | 3 | 5 | 2.681030146 | 2.699973904 | 0.9897552 |
| MAE | 4 | 5 | 4.045828502 | 2.699973904 | 0.501383117 |
| MEDAE | 2 | 3 | 3.340871429 | 2.298098753 | 0.519157427 |
| MEDAE | 2 | 4 | 3.340871429 | 3.644910628 | 0.885358093 |
| MEDAE | 2 | 5 | 3.340871429 | 2.306331942 | 0.523705791 |
| MEDAE | 3 | 4 | 2.298098753 | 3.644910628 | 0.500421055 |
| MEDAE | 3 | 5 | 2.298098753 | 2.306331942 | 0.995549459 |
| MEDAE | 4 | 5 | 3.644910628 | 2.306331942 | 0.503880028 |

*B.2.3 Adam*

*Activation Functions*

| Error Metric | Activation Function i | Activation Function j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | identity | logistic | 24.48079902 | 3.878377003 | 2.22E-07 |
| MAE | identity | tanh | 24.48079902 | 5.801421891 | 2.62E-06 |
| MAE | identity | relu | 24.48079902 | 12.43272562 | 0.005659804 |
| MAE | logistic | tanh | 3.878377003 | 5.801421891 | 8.82E-10 |
| MAE | logistic | relu | 3.878377003 | 12.43272562 | 4.62E-06 |
| MAE | tanh | relu | 5.801421891 | 12.43272562 | 0.000403393 |
| MEDAE | identity | logistic | 23.55389723 | 3.545543675 | 3.84E-07 |
| MEDAE | identity | tanh | 23.55389723 | 5.462041105 | 4.40E-06 |
| MEDAE | identity | relu | 23.55389723 | 7.985455799 | 0.000132035 |
| MEDAE | logistic | tanh | 3.545543675 | 5.462041105 | 3.50E-09 |
| MEDAE | logistic | relu | 3.545543675 | 7.985455799 | 8.04E-05 |
| MEDAE | tanh | relu | 5.462041105 | 7.985455799 | 0.027493137 |

*Spatial Lags*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 10.02654338 | 10.47054031 | 0.837434719 |
| MEDAE | 0 | 1 | 9.55306945 | 8.656085784 | 0.66919061 |

*Temporal Lags*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 12.59776236 | 11.94973591 | 0.808312304 |
| MAE | 2 | 4 | 12.59776236 | 9.24116642 | 0.174435804 |
| MAE | 2 | 5 | 12.59776236 | 6.354184955 | 0.002291553 |
| MAE | 3 | 4 | 11.94973591 | 9.24116642 | 0.325070362 |
| MAE | 3 | 5 | 11.94973591 | 6.354184955 | 0.018695996 |
| MAE | 4 | 5 | 9.24116642 | 6.354184955 | 0.179389911 |
| MEDAE | 2 | 3 | 11.13795349 | 9.883020618 | 0.595549167 |
| MEDAE | 2 | 4 | 11.13795349 | 8.60312028 | 0.286073863 |
| MEDAE | 2 | 5 | 11.13795349 | 4.662745536 | 0.000192937 |
| MEDAE | 3 | 4 | 9.883020618 | 8.60312028 | 0.610628906 |
| MEDAE | 3 | 5 | 9.883020618 | 4.662745536 | 0.00658365 |
| MEDAE | 4 | 5 | 8.60312028 | 4.662745536 | 0.041647448 |

## B.3 Species Analysis SVM

### *B.3.1 Linear*

*Spatial Lags*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 4.987698895 | 6.275510309 | 0.040509883 |
| MEDAE | 0 | 1 | 5.370444751 | 6.766721649 | 0.043513211 |

*Temporal Lags*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 5.502255208 | 4.782759259 | 0.253120689 |
| MAE | 2 | 4 | 5.502255208 | 6.137632075 | 0.532762229 |
| MAE | 2 | 5 | 5.502255208 | 5.63715625 | 0.861892624 |
| MAE | 3 | 4 | 4.782759259 | 6.137632075 | 0.18358271 |
| MAE | 3 | 5 | 4.782759259 | 5.63715625 | 0.268569602 |
| MAE | 4 | 5 | 6.137632075 | 5.63715625 | 0.652031052 |
| MEDAE | 2 | 3 | 5.950546875 | 5.134419753 | 0.245738275 |
| MEDAE | 2 | 4 | 5.950546875 | 6.516849057 | 0.610724801 |
| MEDAE | 2 | 5 | 5.950546875 | 6.164354167 | 0.803503444 |
| MEDAE | 3 | 4 | 5.134419753 | 6.516849057 | 0.213641713 |
| MEDAE | 3 | 5 | 5.134419753 | 6.164354167 | 0.228570487 |
| MEDAE | 4 | 5 | 6.516849057 | 6.164354167 | 0.771065501 |

### *B.3.2 Polynomial*

*Degree*

| Error Metric | Degree i | Degree j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 6.865614865 | 18.34186057 | 0.107481212 |
| MAE | 2 | 4 | 6.865614865 | 9.808818841 | 0.022585635 |
| MAE | 3 | 4 | 18.34186057 | 9.808818841 | 0.234413705 |
| MEDAE | 2 | 3 | 7.170074324 | 18.82694316 | 0.105127039 |
| MEDAE | 2 | 4 | 7.170074324 | 9.770072464 | 0.024302358 |
| MEDAE | 3 | 4 | 18.82694316 | 9.770072464 | 0.209087718 |

*Spatial Lags*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 13.49366119 | 8.094828358 | 0.143730576 |
| MEDAE | 0 | 1 | 13.57039684 | 8.785634328 | 0.198371287 |

*Temporal Lags*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 11.13869301 | 19.03654717 | 0.424486101 |
| MAE | 2 | 4 | 11.13869301 | 8.770367347 | 0.462781468 |
| MAE | 2 | 5 | 11.13869301 | 6.71815625 | 0.157293695 |
| MAE | 3 | 4 | 19.03654717 | 8.770367347 | 0.28188121 |
| MAE | 3 | 5 | 19.03654717 | 6.71815625 | 0.196020086 |
| MAE | 4 | 5 | 8.770367347 | 6.71815625 | 0.180688265 |
| MEDAE | 2 | 3 | 11.54204557 | 19.3404717 | 0.431821116 |
| MEDAE | 2 | 4 | 11.54204557 | 8.715418367 | 0.383473647 |
| MEDAE | 2 | 5 | 11.54204557 | 7.007765625 | 0.162253291 |
| MEDAE | 3 | 4 | 19.3404717 | 8.715418367 | 0.264315354 |
| MEDAE | 3 | 5 | 19.3404717 | 7.007765625 | 0.195871042 |
| MEDAE | 4 | 5 | 8.715418367 | 7.007765625 | 0.219894487 |

*B.3.3 Radial Base Function*

*Spatial Lags*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 6.088 | 9.038934211 | 0.000656458 |
| MEDAE | 0 | 1 | 6.566862385 | 9.897131579 | 0.000432648 |

*Temporal Lags*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 6.047488889 | 7.490729167 | 0.172635538 |
| MAE | 2 | 4 | 6.047488889 | 6.130810345 | 0.944672954 |
| MAE | 2 | 5 | 6.047488889 | 7.90344 | 0.136085822 |
| MAE | 3 | 4 | 7.490729167 | 6.130810345 | 0.257150217 |
| MAE | 3 | 5 | 7.490729167 | 7.90344 | 0.736285296 |
| MAE | 4 | 5 | 6.130810345 | 7.90344 | 0.193810311 |
| MEDAE | 2 | 3 | 6.580588889 | 8.091114583 | 0.200188323 |
| MEDAE | 2 | 4 | 6.580588889 | 6.584327586 | 0.997792079 |
| MEDAE | 2 | 5 | 6.580588889 | 8.65734 | 0.134402161 |
| MEDAE | 3 | 4 | 8.091114583 | 6.584327586 | 0.264170188 |
| MEDAE | 3 | 5 | 8.091114583 | 8.65734 | 0.678186758 |
| MEDAE | 4 | 5 | 6.584327586 | 8.65734 | 0.175795759 |

*B.3.4 Hyperbolic Tangent Function*

*Spatial Lags*

| Error Metric | Spatial Lag i | Spatial Lag j | $\overline{x}_j$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 0 | 1 | 31.31246341 | 9.326964286 | 0.067438756 |
| MEDAE | 0 | 1 | 19.47393293 | 10.32666667 | 0.286897559 |

*Temporal Lags*

| Error Metric | Temporal Lag i | Temporal Lag j | $\overline{x}_i$ | $\overline{x}_j$ | $P(T < -t \cap T > t)$ |
|---|---|---|---|---|---|
| MAE | 2 | 3 | 21.31061538 | 6.959869565 | 0.264758932 |
| MAE | 2 | 4 | 21.31061538 | 53.70370455 | 0.340447002 |
| MAE | 2 | 5 | 21.31061538 | 36.85911765 | 0.628936452 |
| MAE | 3 | 4 | 6.959869565 | 53.70370455 | 0.145445968 |
| MAE | 3 | 5 | 6.959869565 | 36.85911765 | 0.319495273 |
| MAE | 4 | 5 | 53.70370455 | 36.85911765 | 0.693806585 |
| MEDAE | 2 | 3 | 9.379666667 | 7.138956522 | 0.034190541 |
| MEDAE | 2 | 4 | 9.379666667 | 32.22538636 | 0.32043919 |
| MEDAE | 2 | 5 | 9.379666667 | 36.90735294 | 0.358685308 |
| MEDAE | 3 | 4 | 7.138956522 | 32.22538636 | 0.276359907 |
| MEDAE | 3 | 5 | 7.138956522 | 36.90735294 | 0.32195311 |
| MEDAE | 4 | 5 | 32.22538636 | 36.90735294 | 0.899448696 |