Predicting Traffic Counts on the National Road Network

Yunhan Wu, Michael Young, Adam Szpiro

October 15, 2020

Background

- Traffic counts are important input for air pollution models.
- Currently no good nationwide traffic data incorporated into air pollution model.
- Goal: generate a national prediction model for traffic counts.

Data Source

1. Traffic data (points):

   - TrafficMetrix: 24-hour average daily traffic counts for fractions of highways and major roads through the United States
   - Sources: city governments, engineering firms, highway and transportation departments.
   - Format: Each row represents one location (road), could contain up to 5 measurements from different years.

2. Road data (lines):

   - TeleAtlas: complete road map across United States
   - Format: GIS line file with features, such as CFCC codes.

Description for Traffic Data

- Time: Measurements are taken from 1950s to 2010s, only consider counts measured after 2000[1].
- Type: Only consider AADT (Average Annual Daily Traffic), ADT (Average Daily Traffic, not been seasonally adjusted)[2].
- Multiple measurements: use median
- Size: about 700,000 to 800,000 locations (roads).

[1] Traffic counts are consistent after 2000, based on analysis of locations with multiple measurements.

[2] Not consider other types such as AAWDT (Average Annual Weekday Traffic).

Description for Road Data

There are about 33 million roads across US. The most important feature we want to extract from road data is CFCC codes, or road types.

- A1: Primary Highway With Limited Access.
- A2: Primary Road Without Limited Access. It consists mainly of US highways, but may include some state highways and county highways that connect cities and larger towns.
- A3: Secondary and Connecting Road, includes mostly state highways, but may include some county highways that connect smaller towns, subdivisions, and neighborhoods.
- A4: Local, Neighborhood, and Rural Road. This category contains a broad range of roads.

Merge Traffic and Road data

- The locations for traffic data are not accurately recorded. There might be ambiguity about which road the record traffic count belongs to. Also, road type information is not contained in the traffic data.
- To assign each traffic count to the correct road, we implement a deterministic algorithm that combines geographical distance between coordinates and textual distance of road names.
- Traffic counts without an assignment are excluded [3]. Information about road type (CFCC codes) is also attached to each traffic record.
- In the end, there are 27,591 A1, 74,605 A2, 313,493 A3 and 381,053 A4 roads. Note that the resulted records are point data.

[3] including identified service roads and ramps

Merge Algorithm Part I

1. If the traffic point doesn't contain a street name, merge it to the nearest road segment if their distance is < 0.0001 degree of latitude or longitude ($\sim$ 10 meters), If no such road segment exists, label the traffic point as no match.

2. For each traffic point containing street name, make a list of road segments within a buffer of < 0.0014 degree of latitude or longitude ($\sim$ 0.1 miles).

3. If the levenshtein distance [4] between the street name of the traffic count and the road name [5] of the road segment is ≤ 3, match the traffic point to the road segment. If there exist multiple qualified road segments, merge to the nearest one. If there is no matching road segment based on names, record the closest road and label it as a possible match.

---

[4] minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other

[5] or potential alternative names

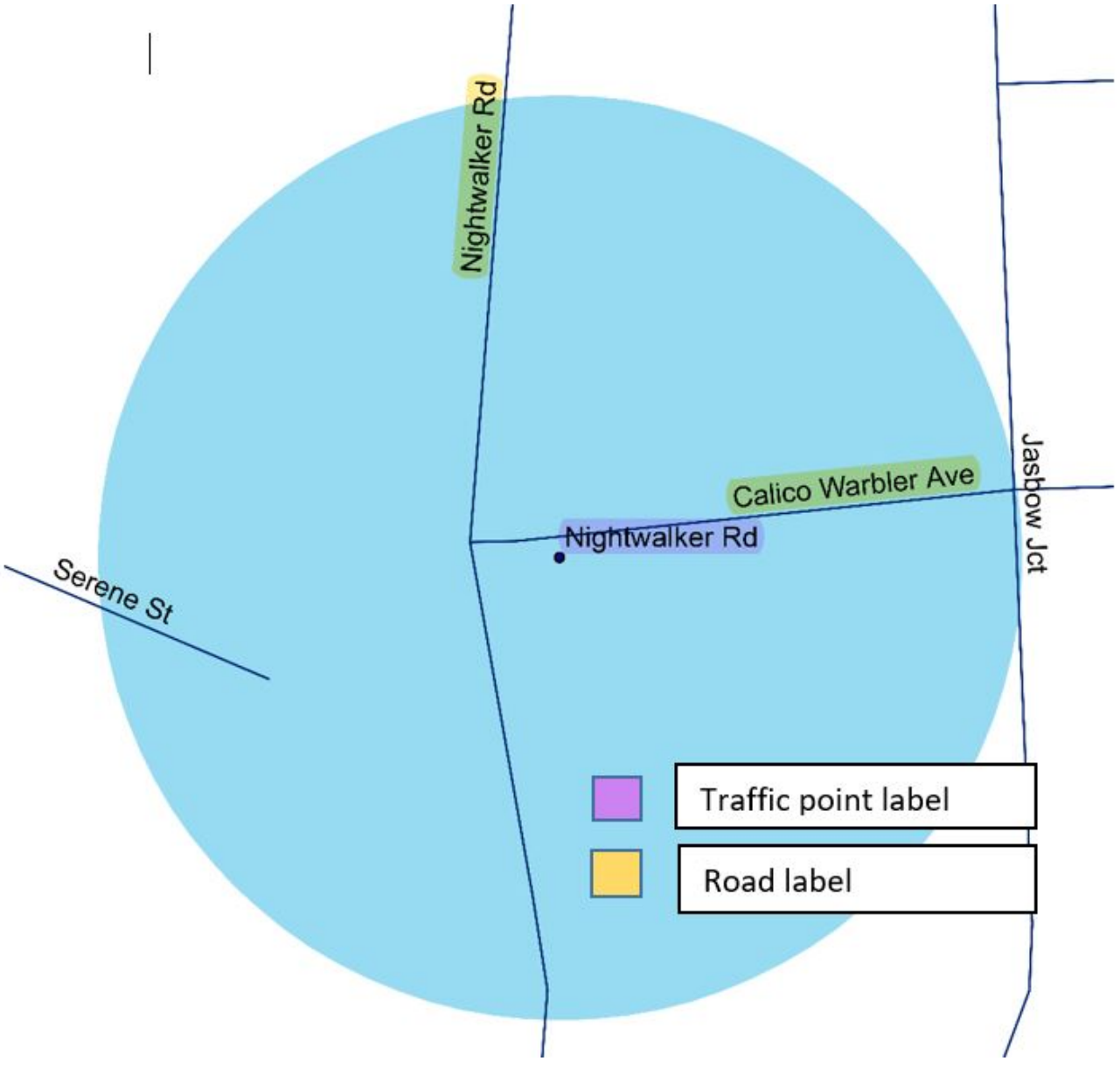
Merge Algorithm Part II

- Within the subset for possible matches, check for systematic naming mismatches. Investigate cases where there are at least three pairs of traffic points and road segments share the same pair of names.

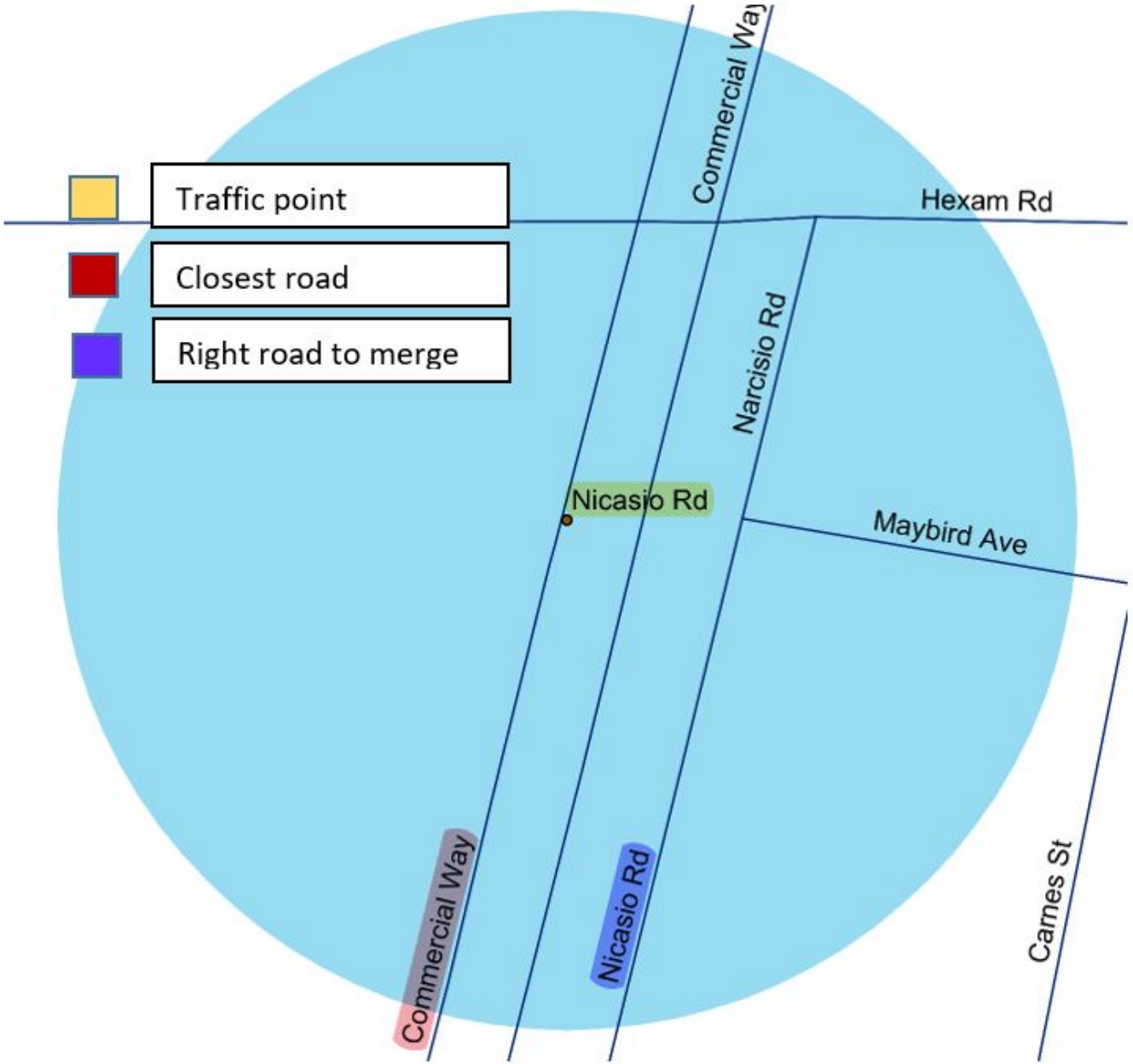
- For example, there might be three traffic points all named 'County Highway' and their corresponding closest road segments share the same name 'Cty Hwy'. Under such case, we will assume 'Cty Hwy' is an alternative name for 'County Highway' and label the possible matches as real matches.

- For the rest of possible matches, label them as no match.

Merge Example I



Merge Example II



Ramp Exclusion for A1

1. If the matched road segment is named 'ramp' or 'Service Road', we remove them from the matched dataset.

2. Fit linear regression model for A1 roads with log traffic counts as outcome, selected land use, ndvi and population density covariates as mean structure. Calculate residual for each observation.

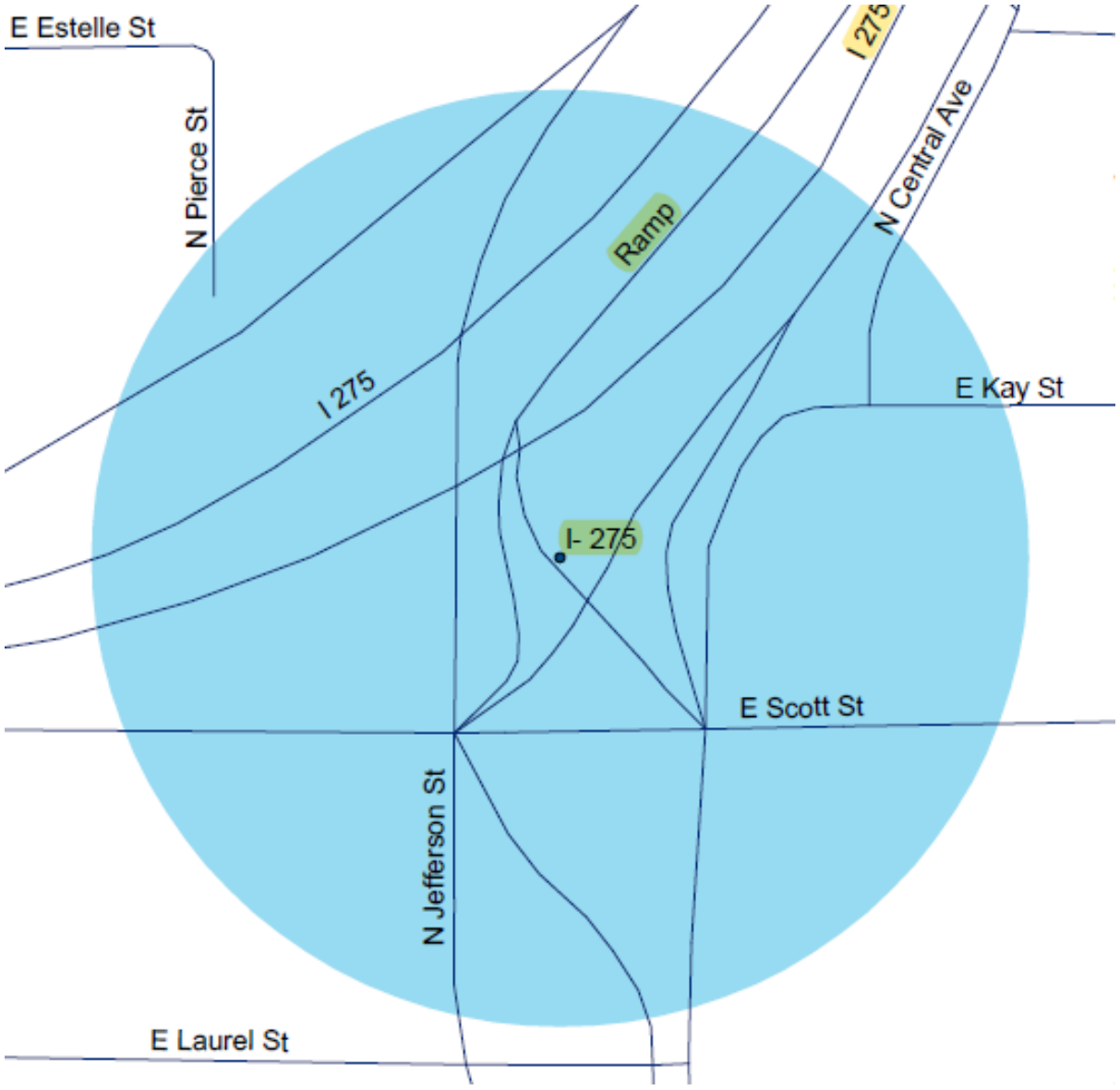
3. First, label observations with residuals < -4 as ramps and exclude them.

4. Find pairs of A1 traffic points within approximately 0.1 mile [6].

5. Exclude the lower one in the pair if its residual < -2.5 and the difference in log traffic counts of the pair > 1.

[6] 0.0015 degree of latitude or longitude

Ramp Example



Geographical Covariates

- Based on the corrected locations, we calculate geographical covariates for each traffic record.
- Major categories of covariates include distance to airports/coastline/railroads/closest A1/A2/A3 roads; population density; land use; impervious surface; ndvi; truck routes, etc. Each covariate comes with different versions with various buffers.
- A representation of cleaned data:

GEOID	CFCC_agre	LINE_LONGI	LINE_LATIT	median_traffic	pop_s00500
TRAFIL0000139923	A1	-87.91917	41.88565	133800	1387
TRAFIL0000139993	A1	-87.85584	41.76343	147500	107
TRAFIL0000140045	A1	-87.58007	41.79729	88500	4691

Important Decisions for Modeling

- Road types: The four road types have different spatial structures[7] and different relationship with covariates. Thus, we decide to model them separately.
- Correlation structure: Network model is ideal for studying the spatial proximity of traffic. However, the road data don't contain direction of traffic flow[8]. Besides, computational complexity will be an issue. Thus, we turn to spatial models based on point data.
- Covariates: In total, there are $\sim$ 800 covariates. Hand pick several [9] vs. dimension reduction using PCA.
- Scale of traffic counts: natural vs. log.

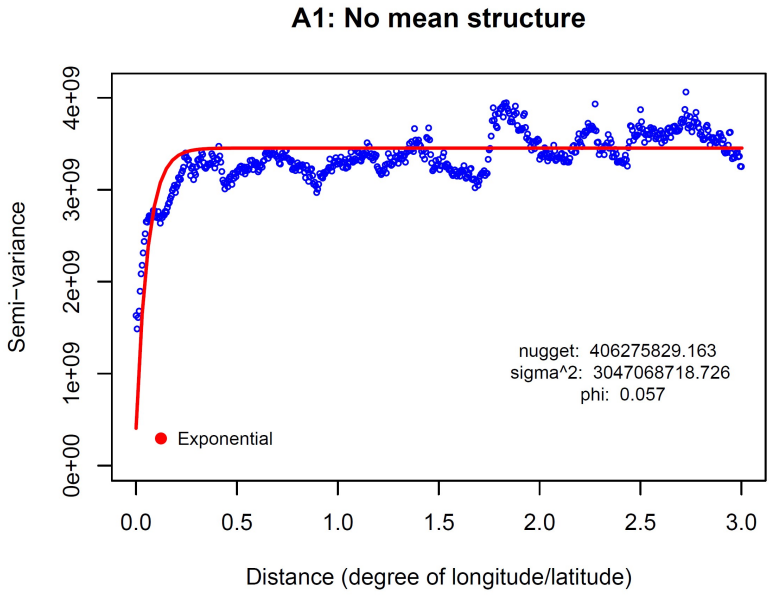
---

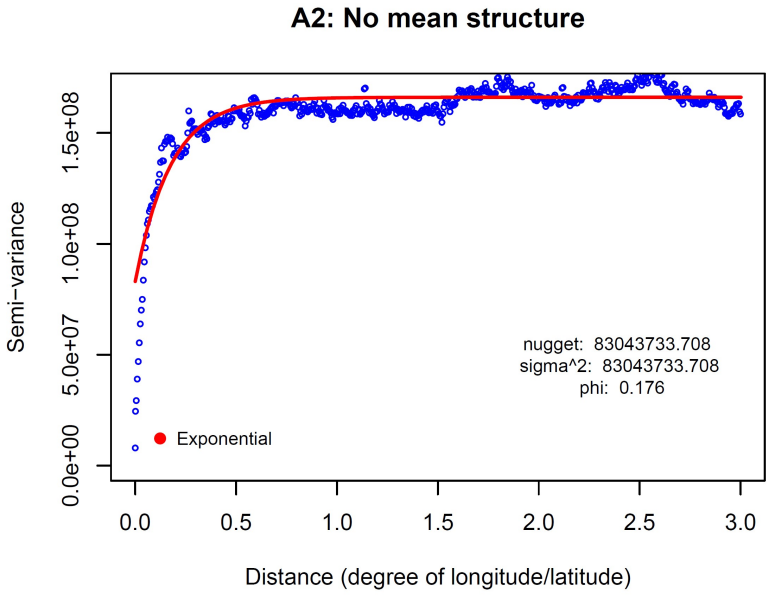
[7] semi-variograms next page

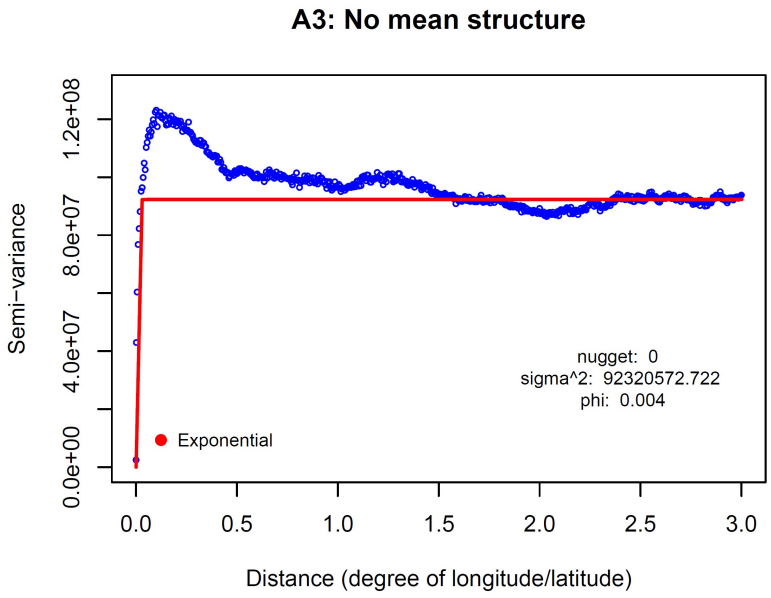
[8] can't tell whether two roads actually intersect from the line file

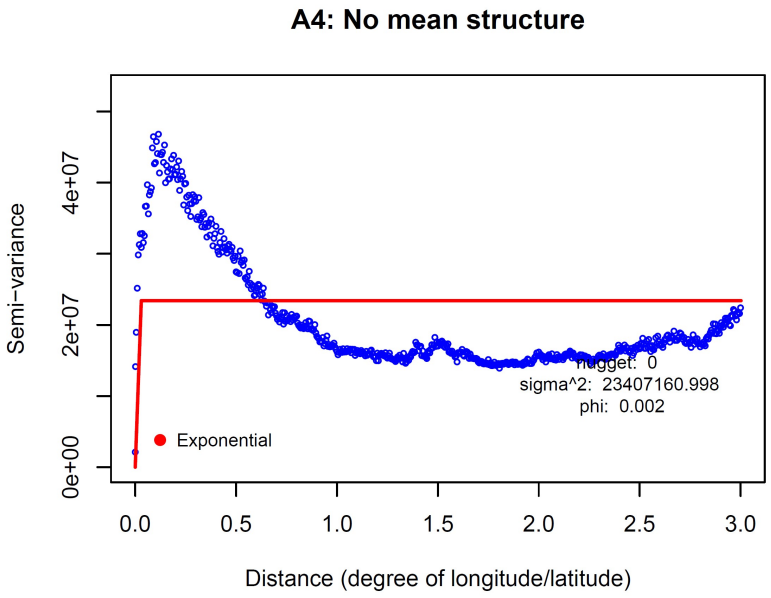
[9] population density, median annual ndvi and low intensity development

Variograms of Natural Scale Traffic Counts

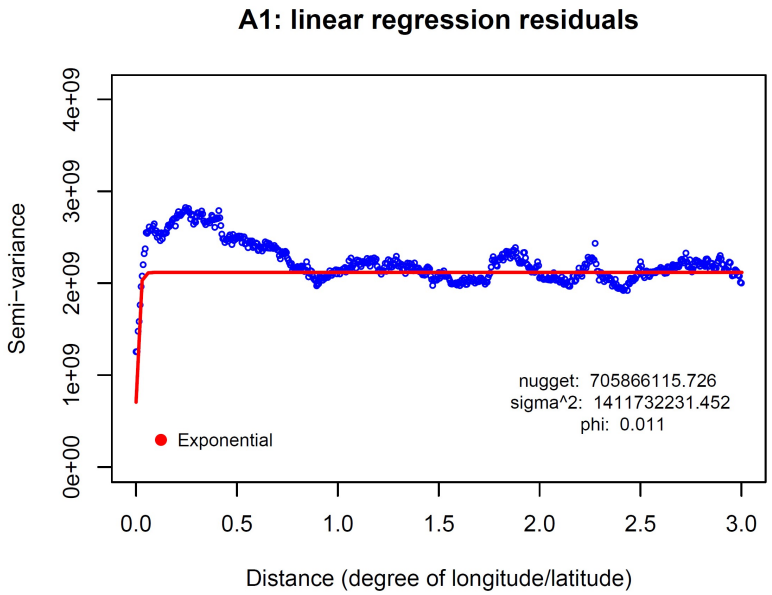


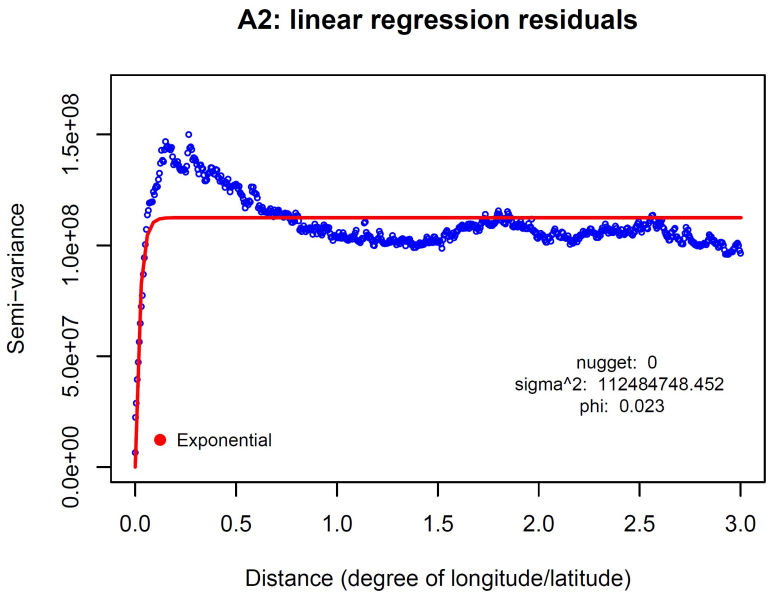


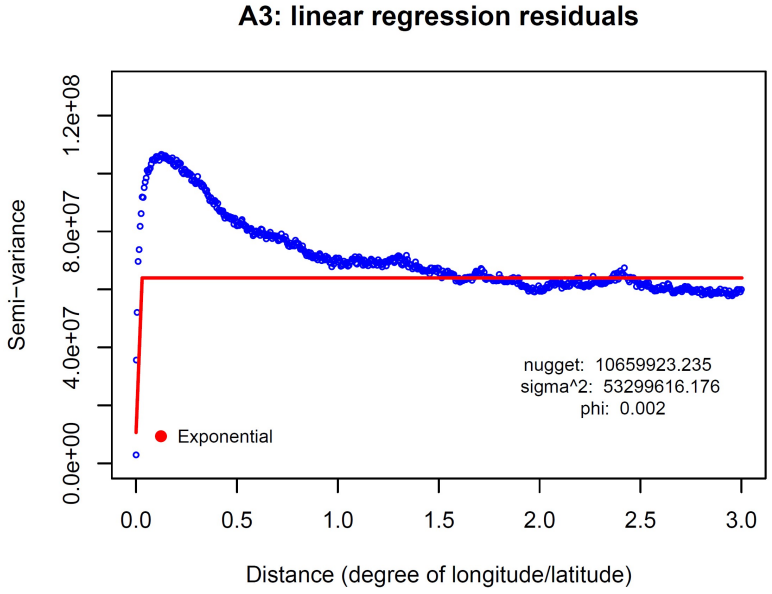


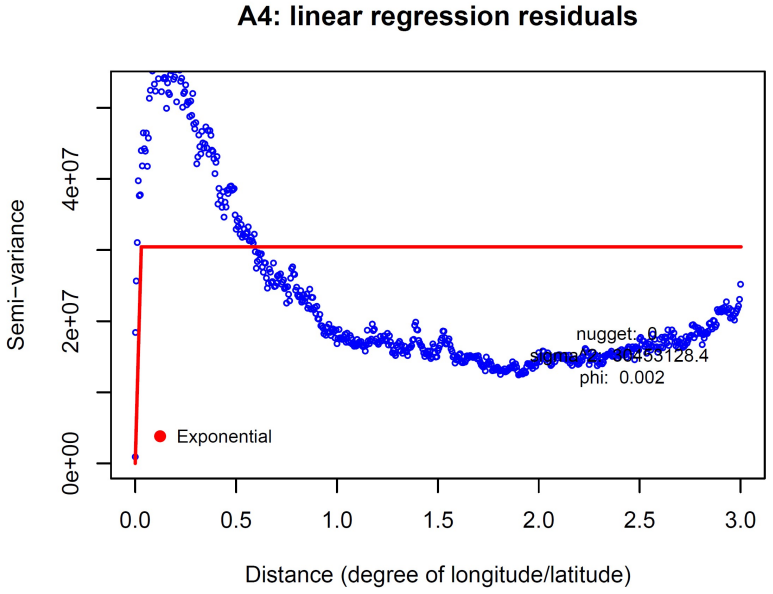


Variograms of Linear Model Residuals[10]



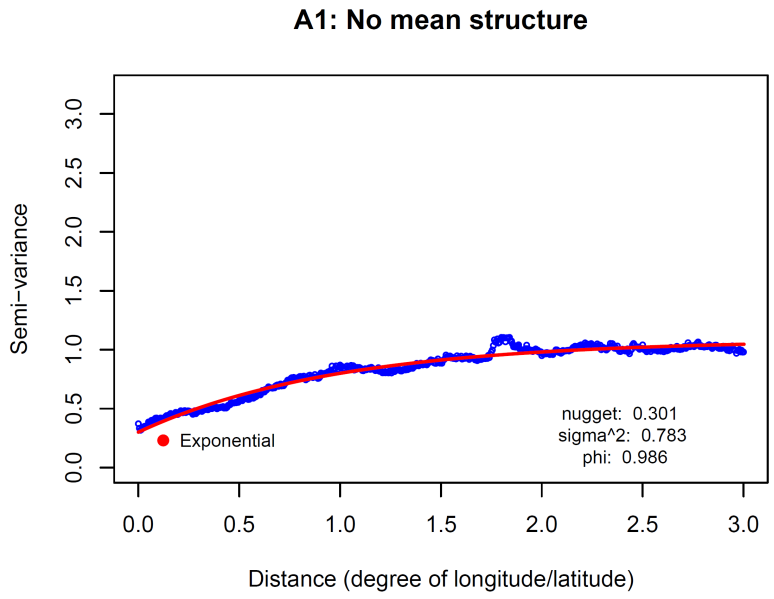


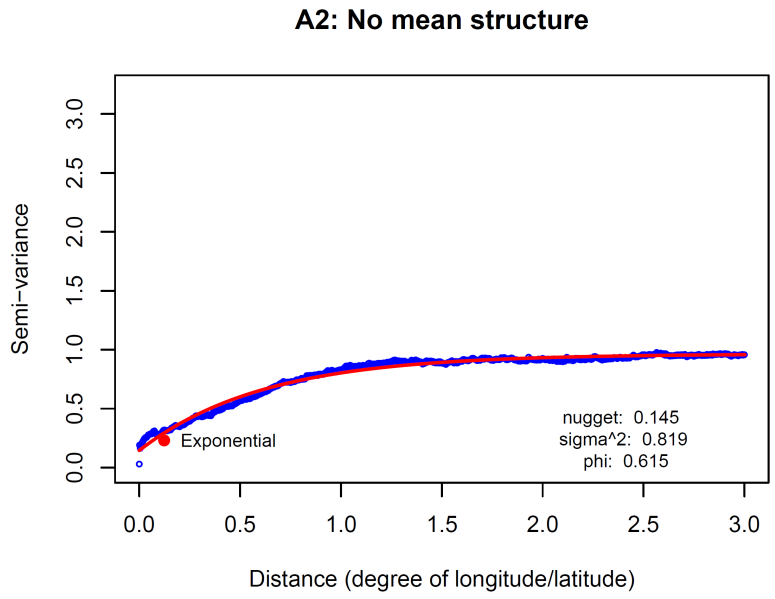


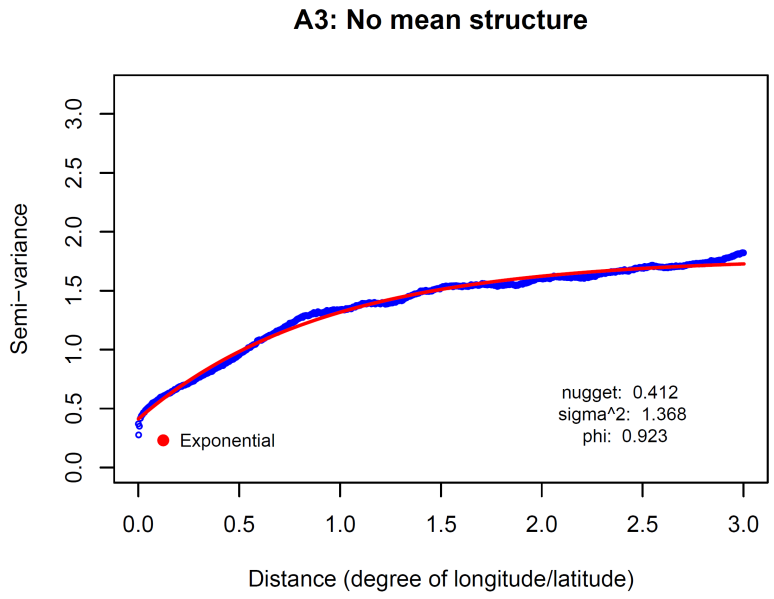


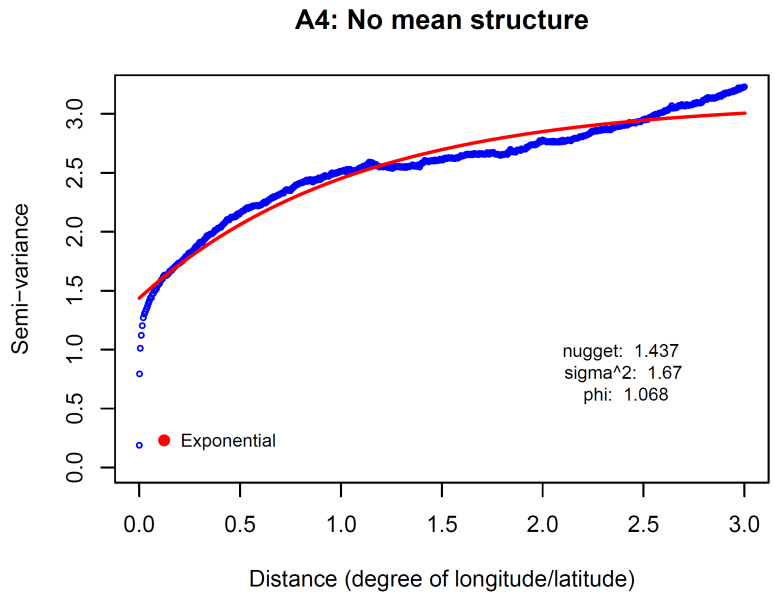
[10]Natural scale traffic counts with selected covariates

Variograms of Log Scale Traffic Counts

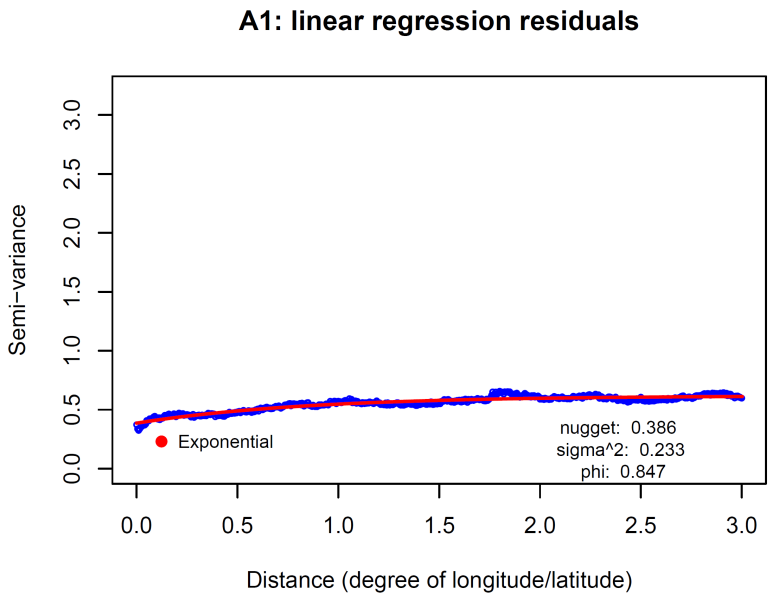


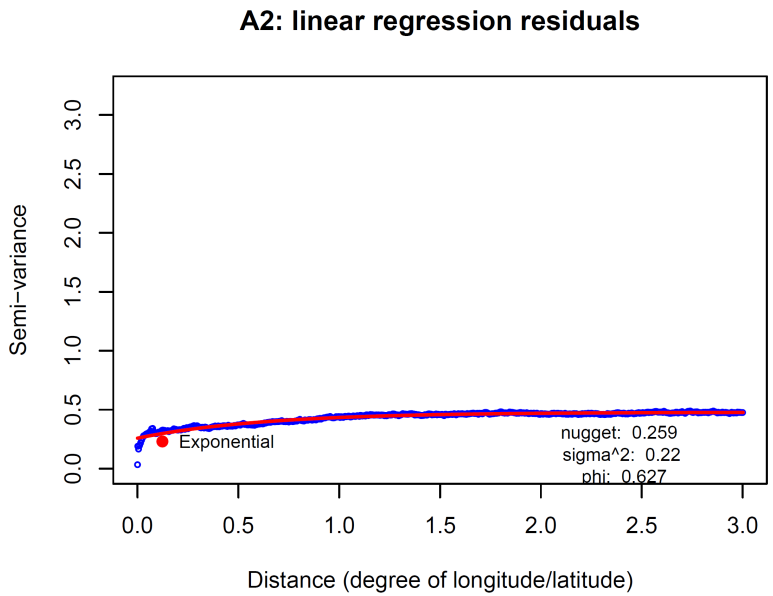


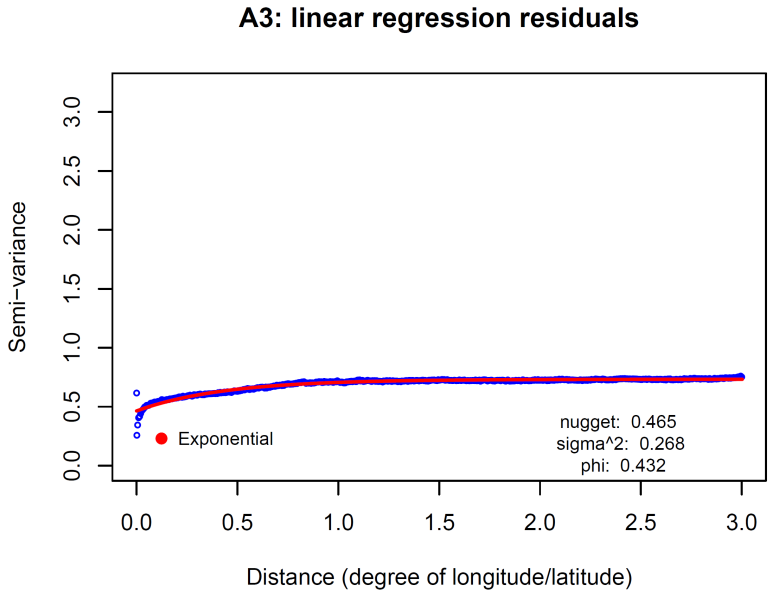


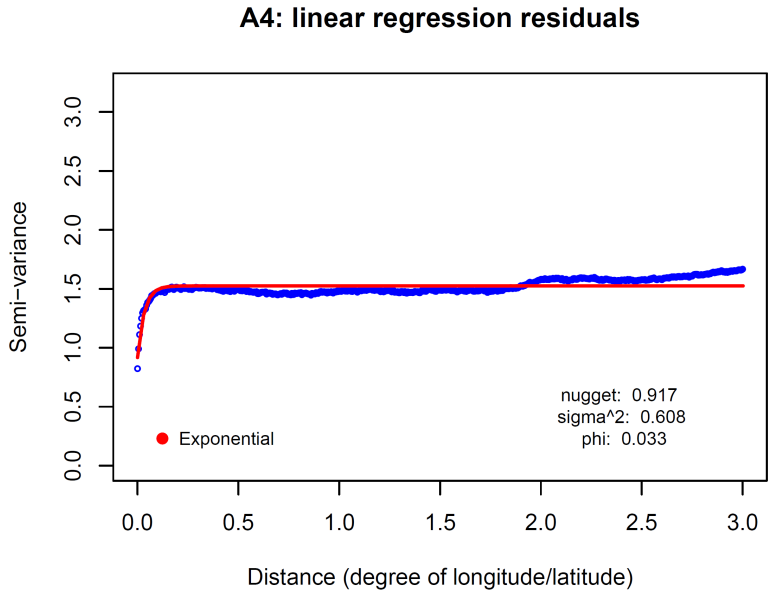


Variograms of Linear Model Residuals[11]









[11] Log traffic counts with selected covariates

Model Choice

We seek to fit spatial model to utilize the spatial structure of traffic counts and incorporate covariates as mean structure.

- Linear regression: Covariates as mean structure, not using spatial information. Used as reference.
- Universal kriging: Standard approach but unfeasible because of $\mathcal{O}(N^3)$ computational complexity. Thus, we did not fit any universal kriging model.
- GAM: Thin plate regression spline (TPRS), fitted using mgcv package in R.
- spNNGP: Nearest Neighbor Gaussian Process models, developed by Finley, A.O., Datta & A., Banerjee S.. It demonstrated highly competitive predictive performance and computation efficiency in Heaton *et al.* (2019) paper among other models for large spatial data.

Thin Plate Regression Spline

We used thin plate regression spline (TPRS) to predict the traffic count. To be specific, we adopt the following semi-parametric representation for the mean structure of each observation:

$$\mathbb{E}(Y_i) = \beta_0 + z_i\beta + f(x_{i1}, x_{i2})$$

where z_i are the covariates and $f(x_{i1}, x_{i2})$ is a smooth function of coordinates, i.e. (longitude,latitude).

Note that this model is unpenalized [12]. We pre-specified the degree of freedom k for the basis functions to construct $f(x_{i1}, x_{i2})$. We consider a sequence of choice of k and compare model performance. Note that we fit separate models for each road type. We consider k in $(15, 50, 150, 500, 1500)$.

[12] Penalized model will require $\mathcal{O}(kN^2)$, so for large k, it is not practical

spNNGP

Under the setting of Gaussian process

$$Y \sim N(X\beta, \Sigma(\phi) + \tau^2\mathbf{I})$$

where $\Sigma(\phi)$ is the spatial covariance function, here we use exponential covariance function and $\tau^2\mathbf{I}$ are iid errors.

spNNGP adopts a Bayesian framework. But instead of using a GP prior for the spatial random effects, it uses a Nearest Neighbor Gaussian Process prior.

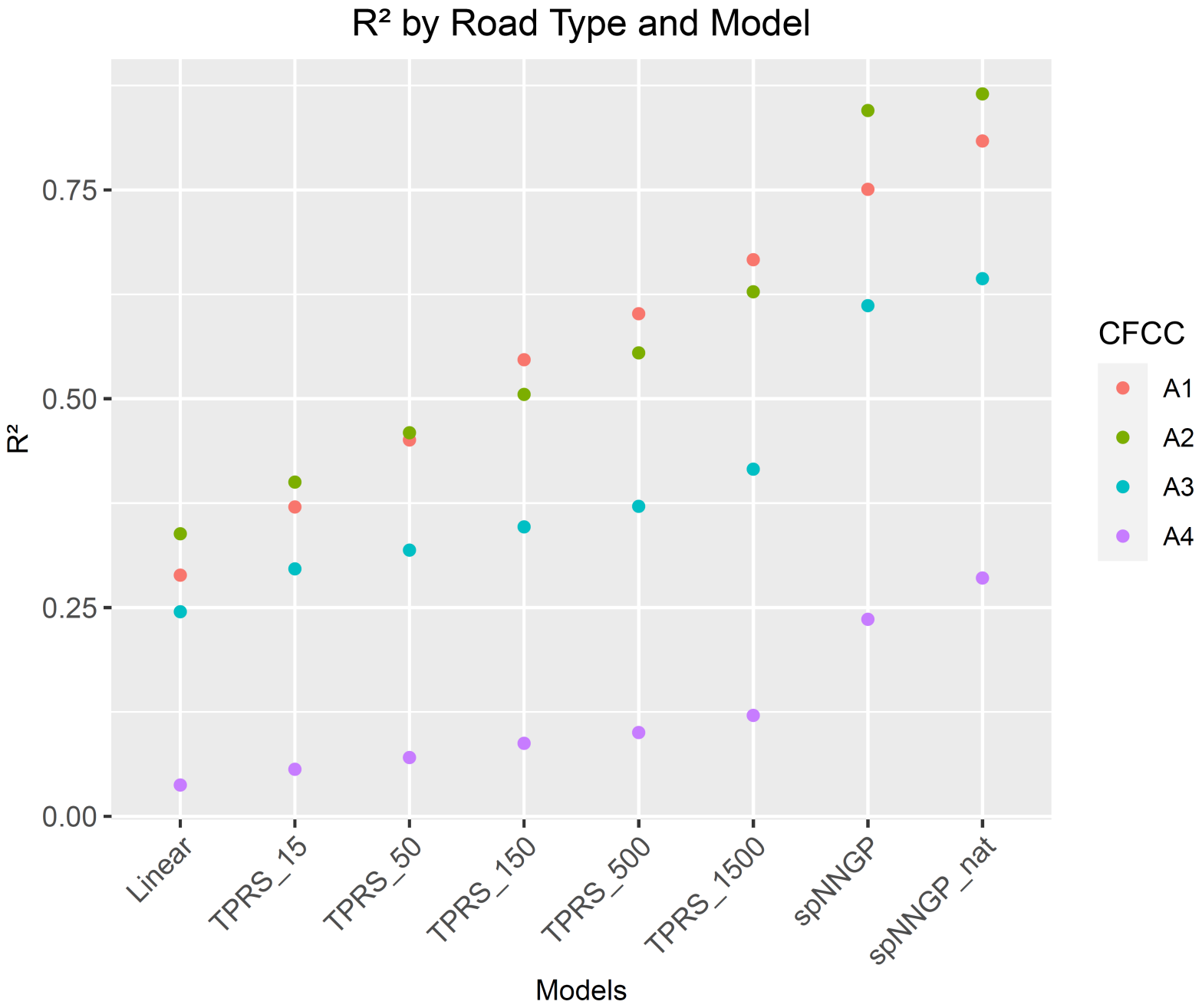
The NNGP approximation $\widetilde{\Sigma}(\phi)$ to $\Sigma(\phi)$ ensures that $\widetilde{\Sigma}(\phi)^{-1}$ is sparse to achieve computational efficiency.

Another feature for spNNGP is to get rid of MCMC. Instead, we prespecify a grid of ϕ and the conjugate NNGP model under spNNGP package will use cross-validation to find an estimate $\hat{\phi}$.

Results for Models with Same Mean Structure

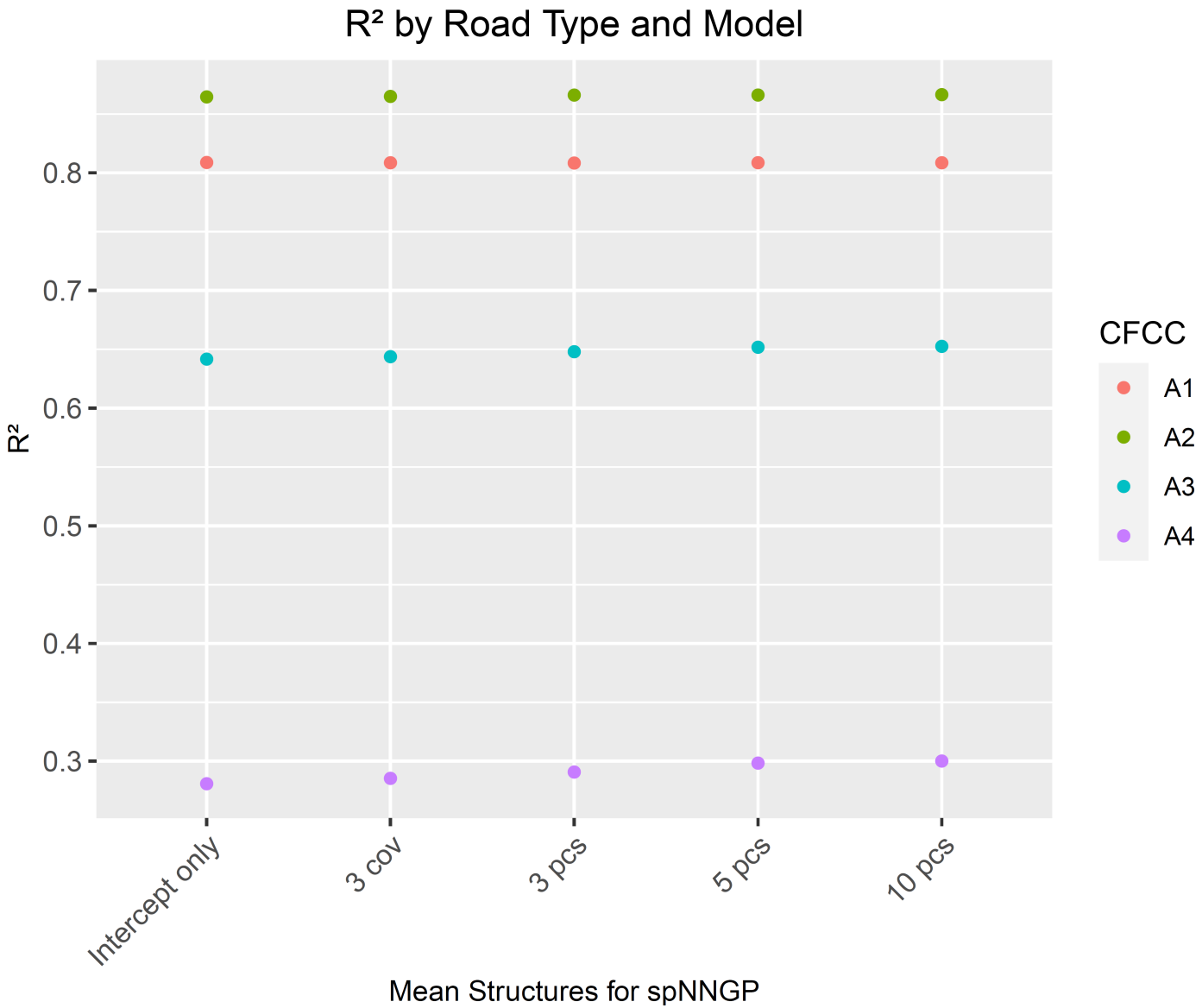
- We first compare different model approaches with the same mean structure (3 selected covariates: population density, median annual ndvi and low intensity development).
- The models we consider are: linear model, TPRS with various numbers of basis function (df=15, 50, 150, 500, 1500), two spNNGP models. All the models except for the last spNNGP are fitted at log scale, i.e. outcomes are log scaled traffic counts. The last model spNNGP_nat are fitted at natural scale.
- R^2 s are calculated based on predicting natural scale traffic counts.
- For each CFCC category, we conduct 10 fold CV to assess the predictive performance.

Results for Models with Same Mean Structure



spNNGP Models with Different Mean Structures

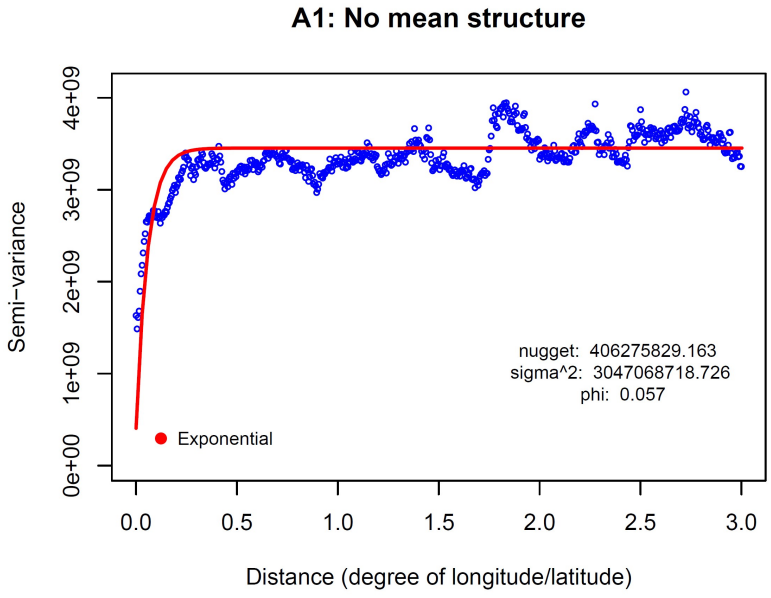
We fit spNNGP models using natural scale traffic counts and compare different mean structures: 0, 3, 5, 10 PCs and 3 selected covariates.

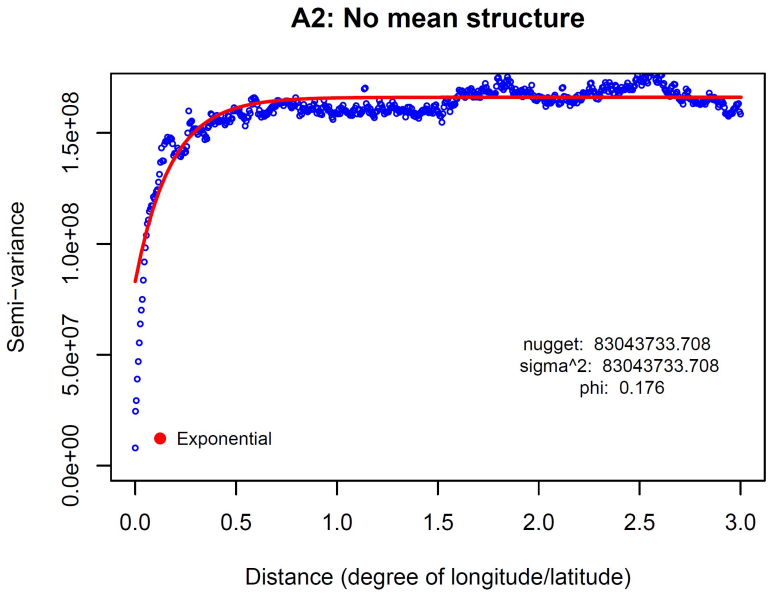


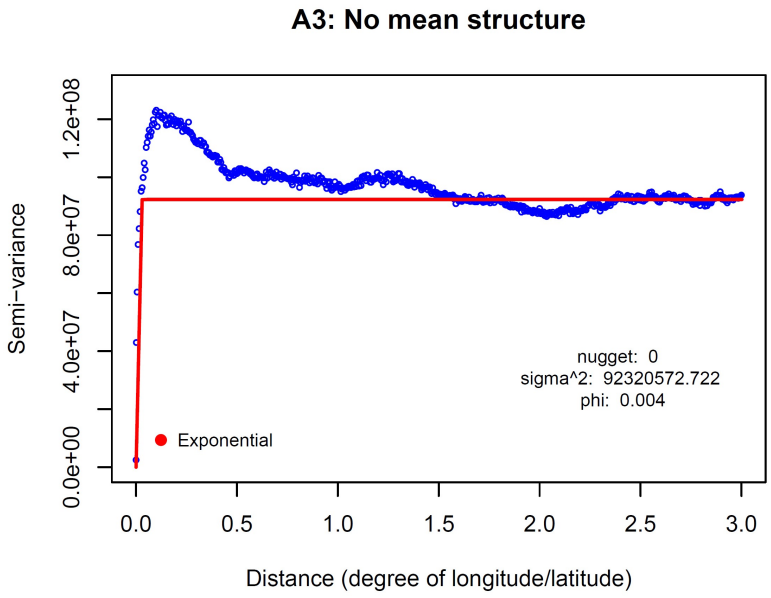
Comparison of variograms

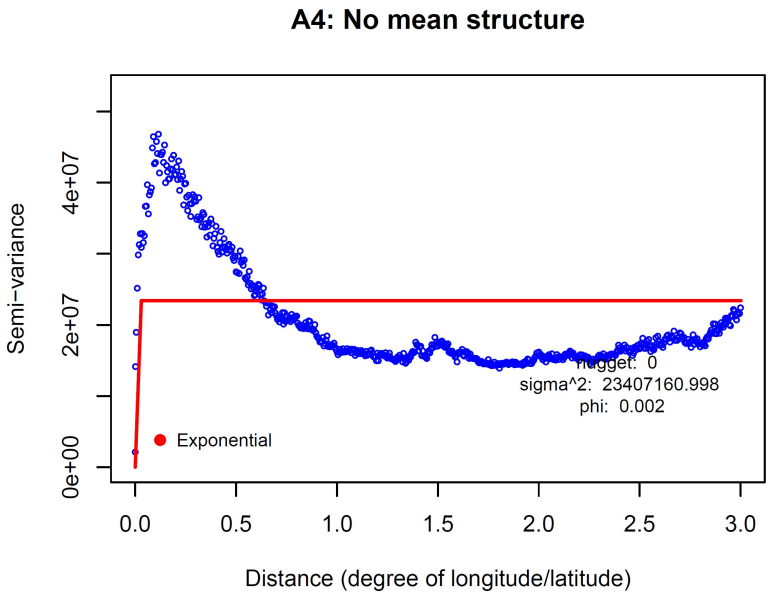
- We compare variograms for raw traffic counts, residuals from linear regression and residuals from spNNGP model.
- Traffic counts and residuals are all at natural scale. Mean structures of the linear regression and spNNGP model for the variograms are the same: three selected covariates.
- We expect to see sills for the variograms to drop as we model spatial patterns at greater detail.

Variograms of Natural Scale Traffic Counts

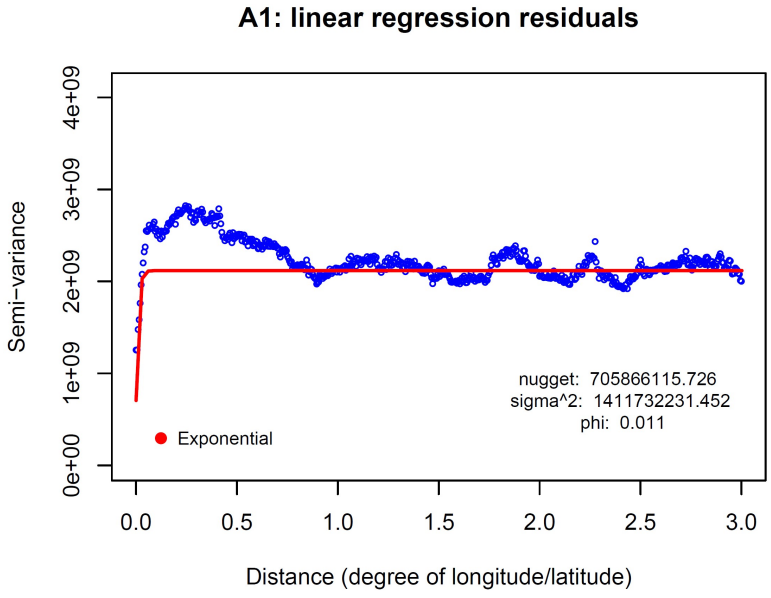


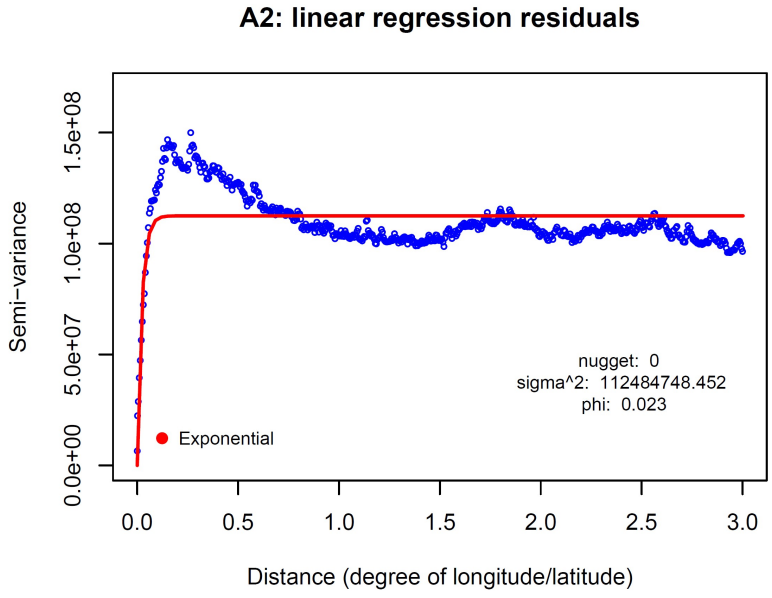


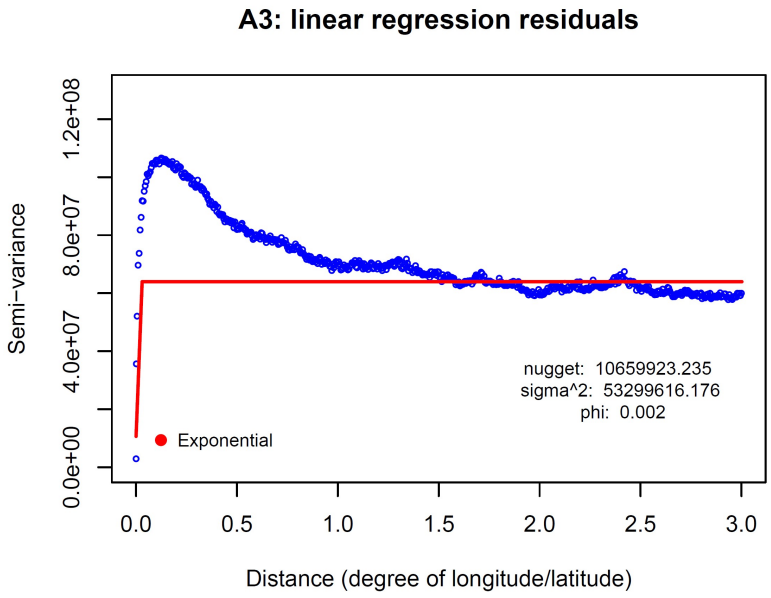


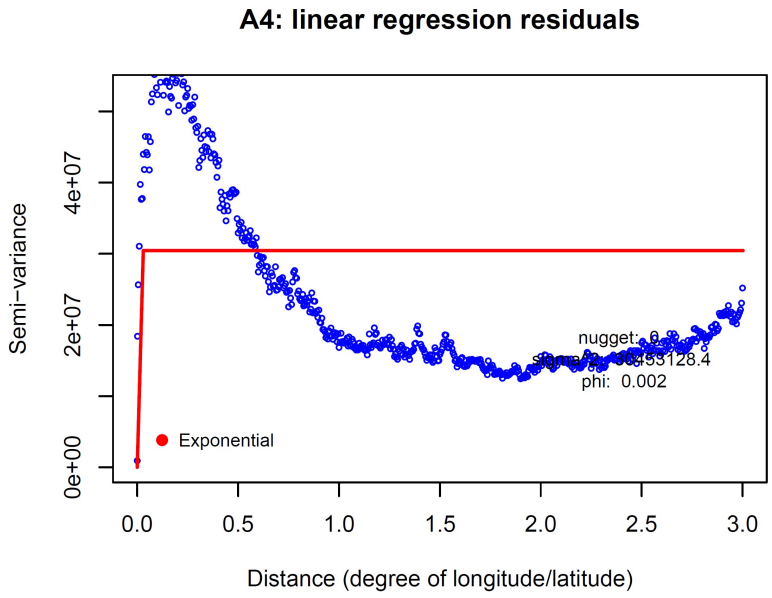


Variograms of Linear Model Residuals

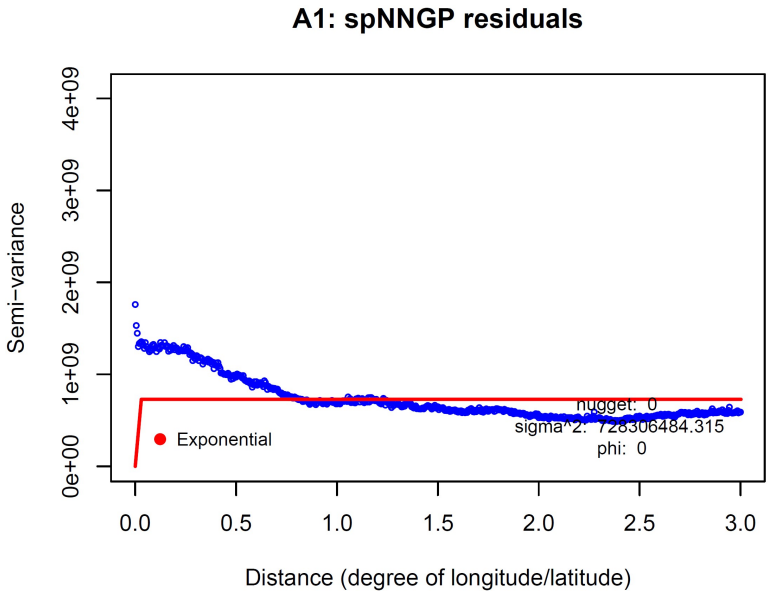


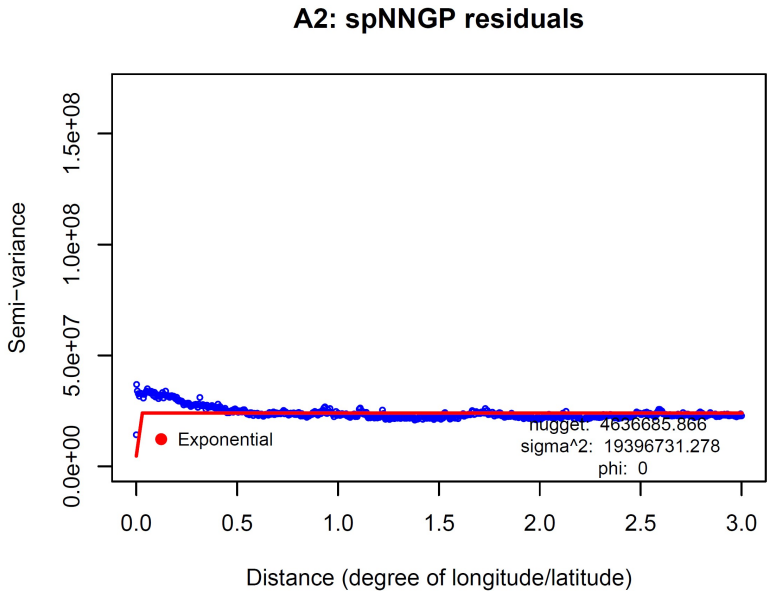


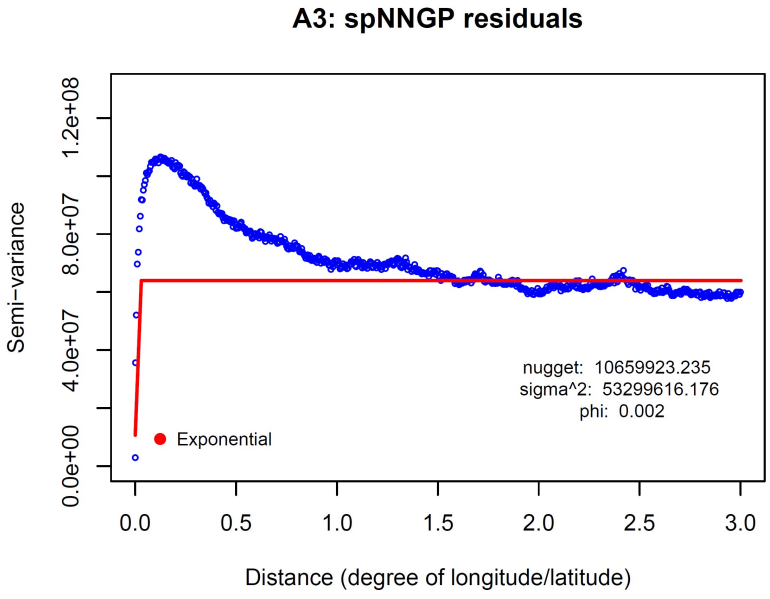


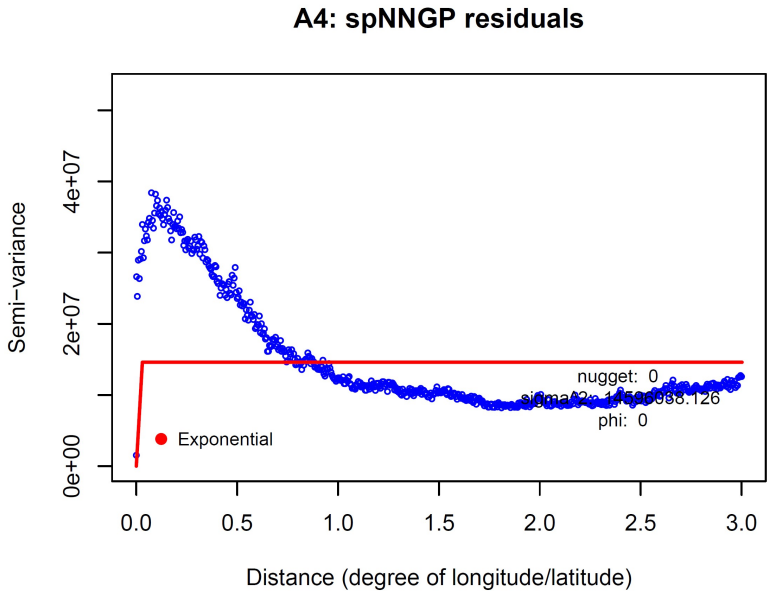


Variograms of spNNGP Residuals

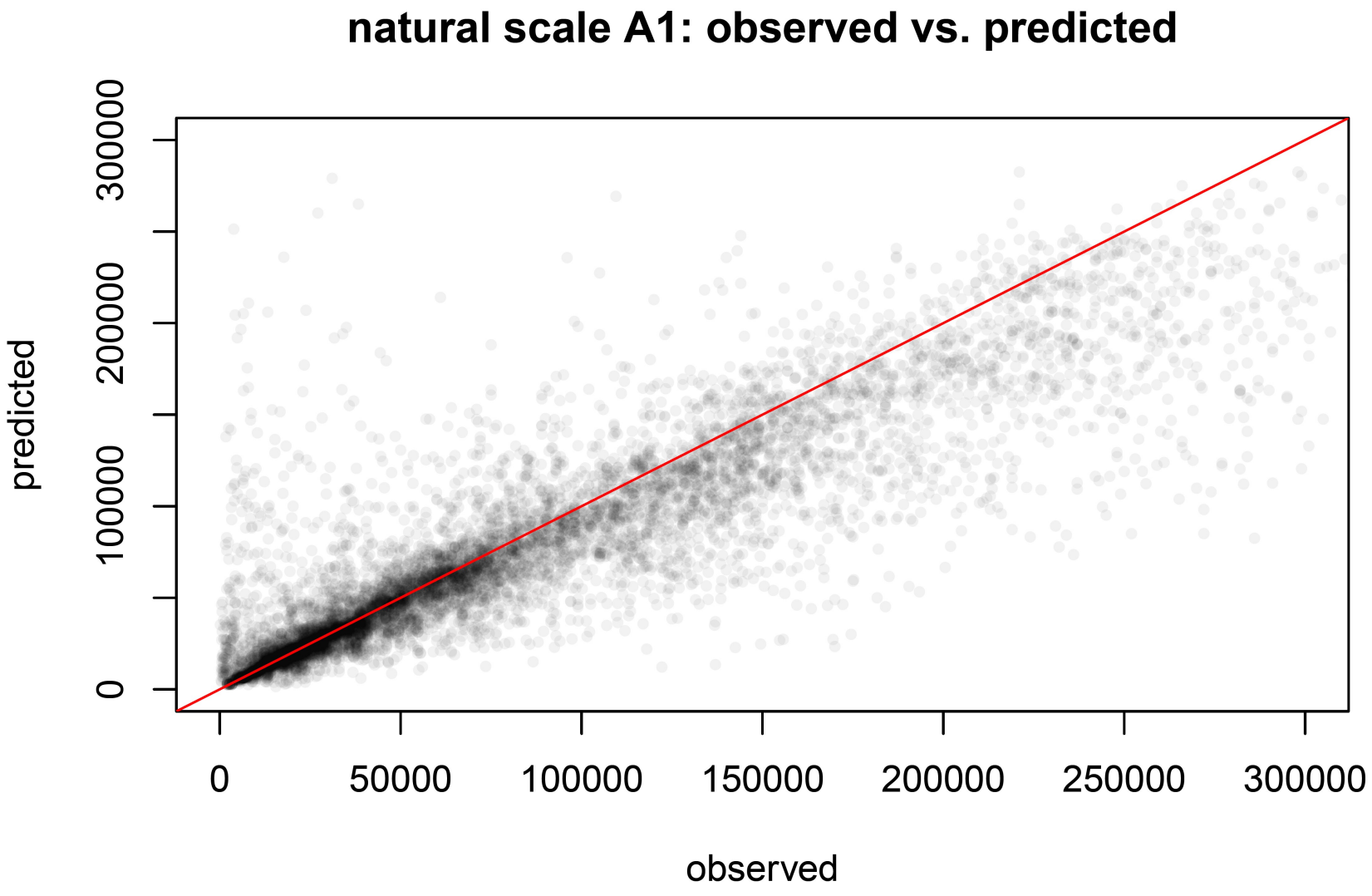




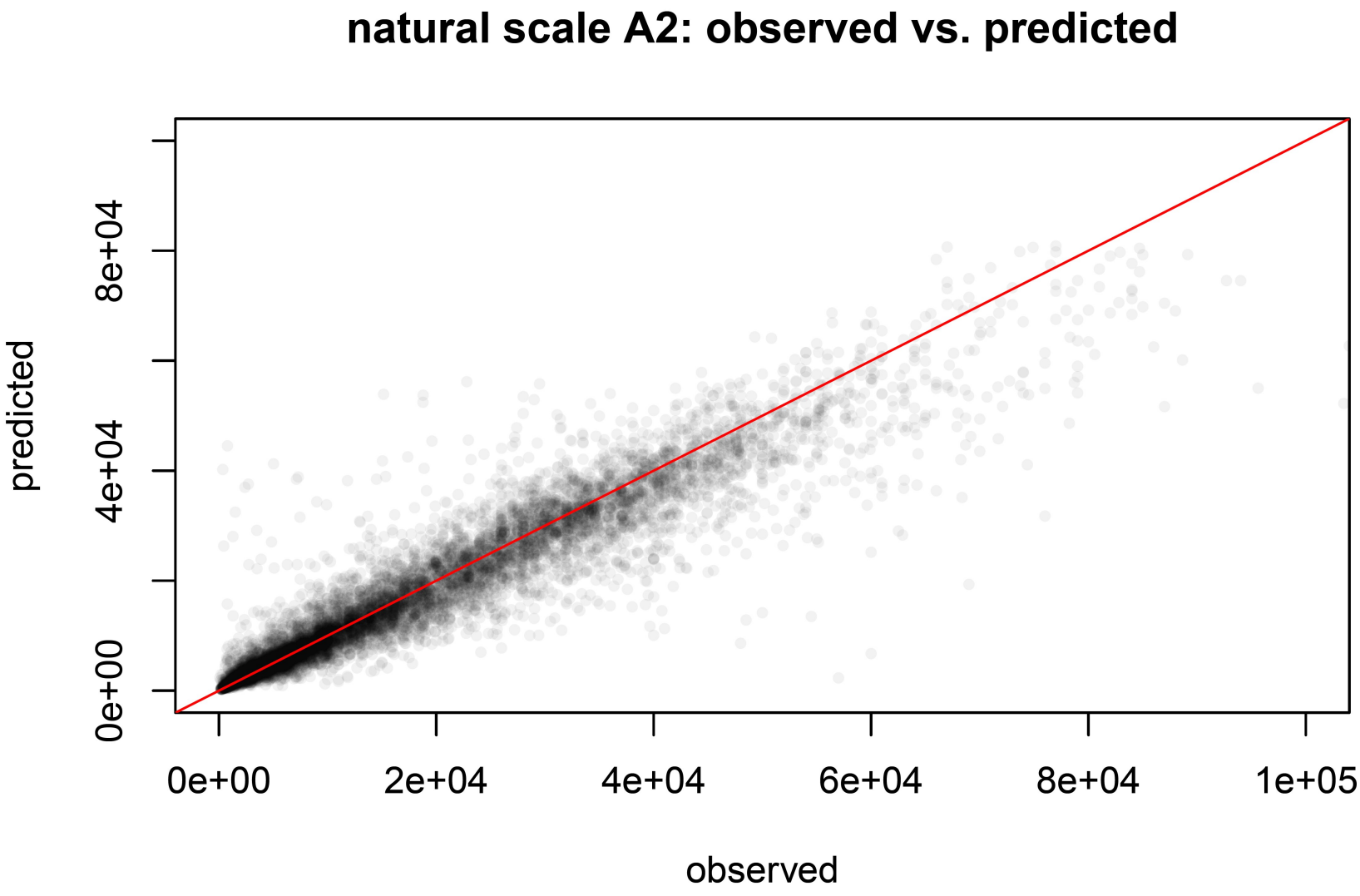




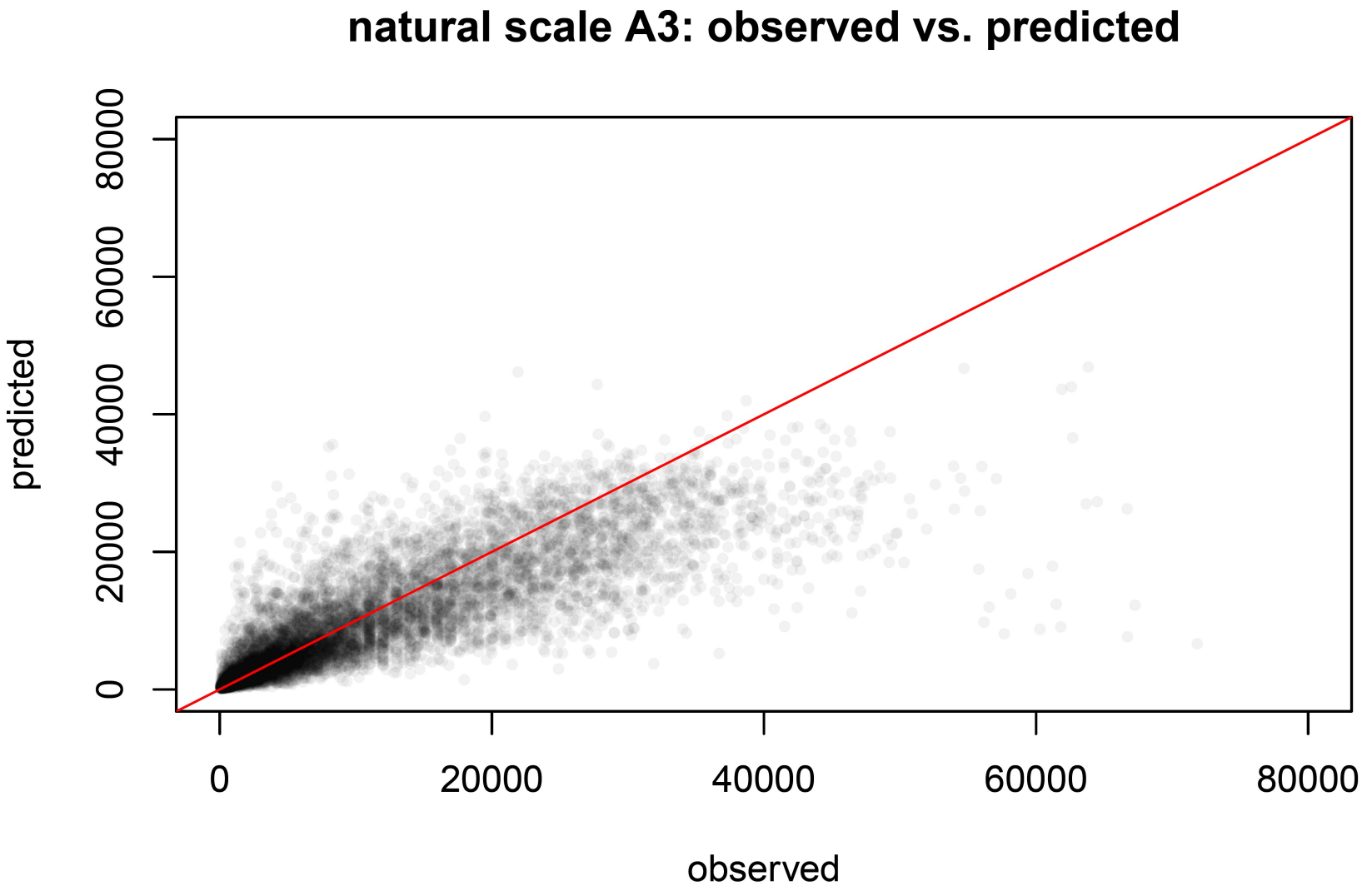
Model Diagnostics: Observed vs. Fitted



Model Diagnostics: Observed vs. Fitted

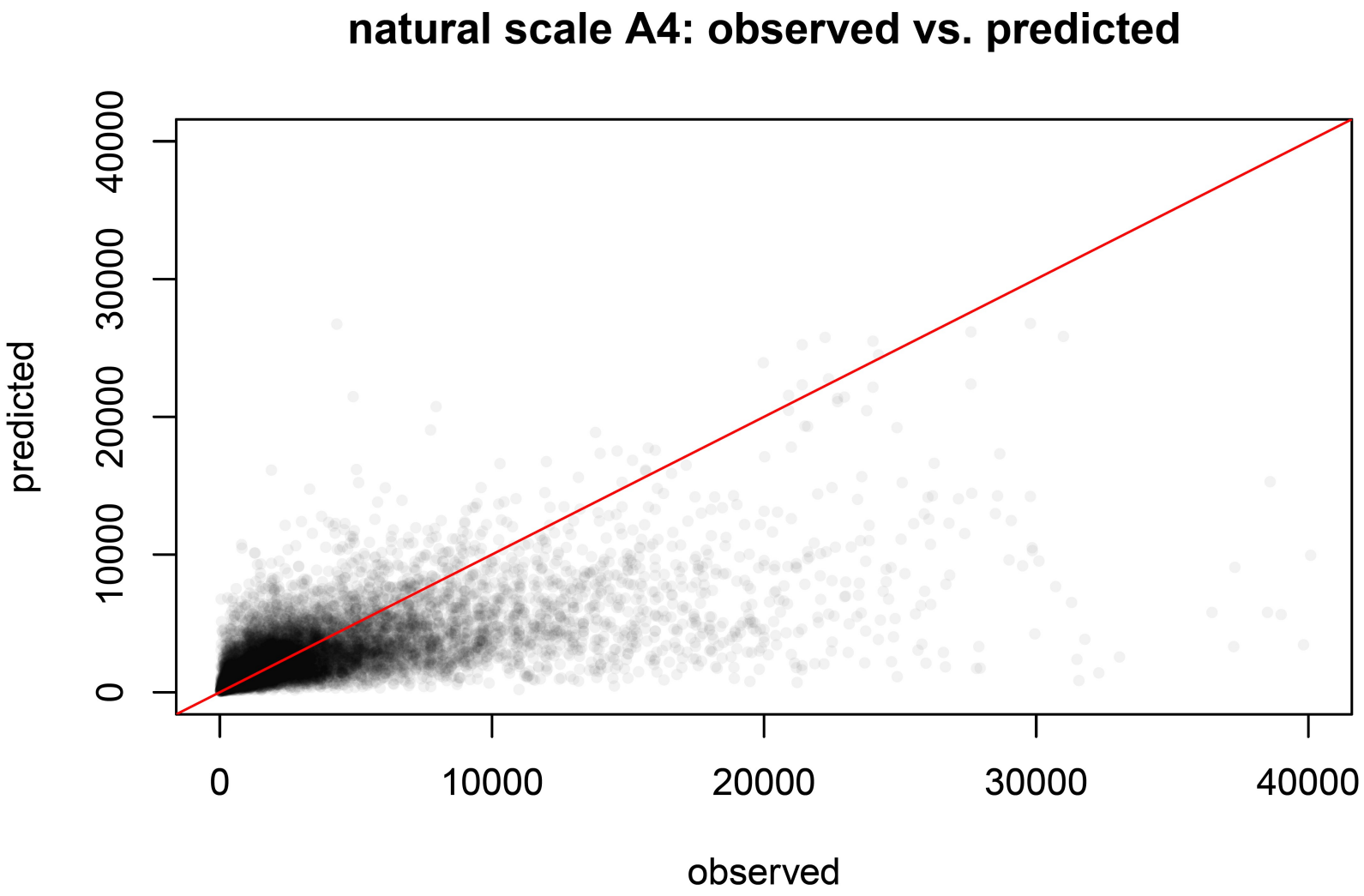


Model Diagnostics: Observed vs. Fitted



Model Diagnostics: Observed vs. Fitted

Issue with A4 roads: underpredicting observations with high traffic counts.

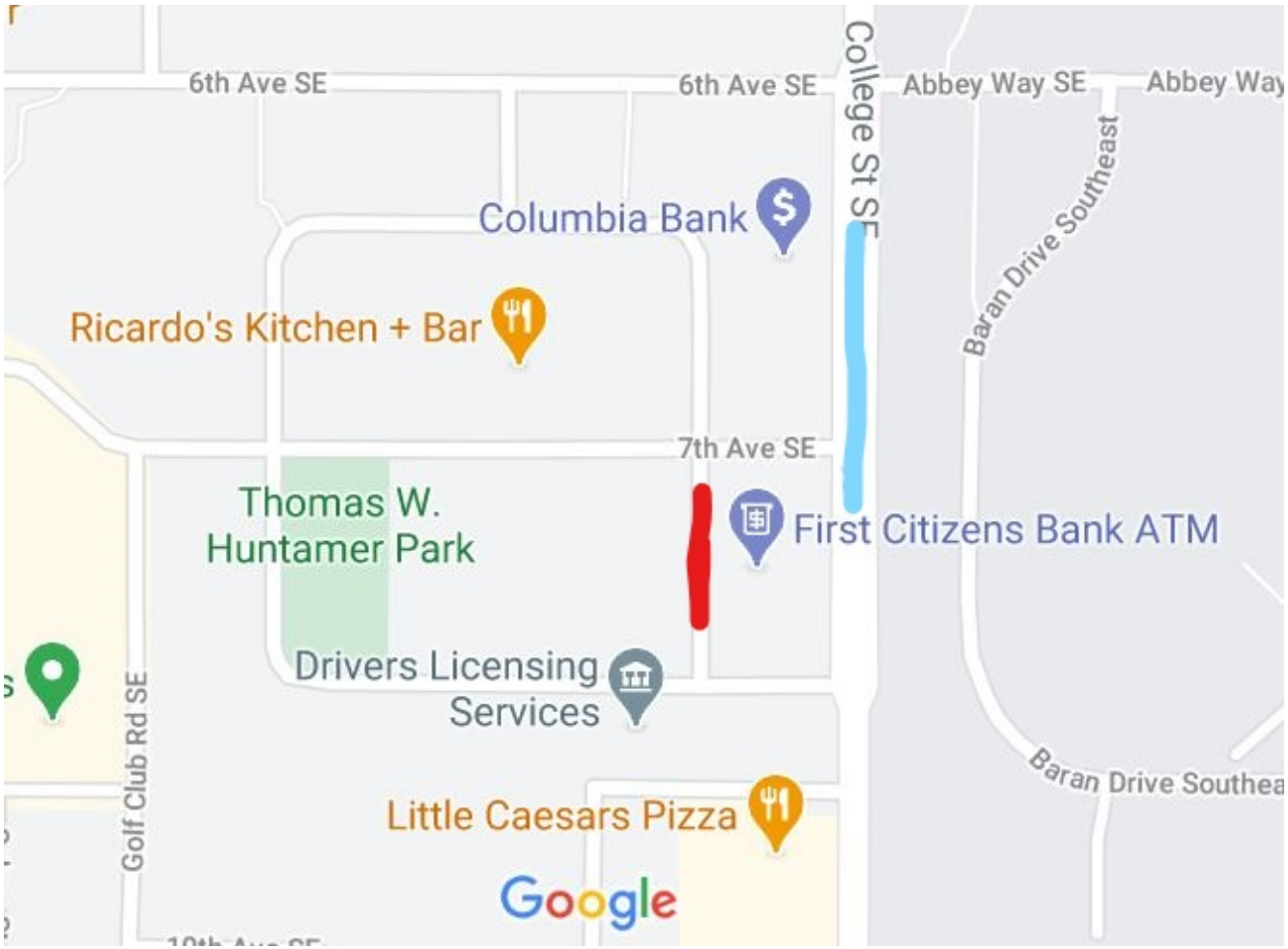


A4 Examples

- The heterogeneity of A4 roads make it hard to predict. In the following slides, we present pairs of A4 roads that are very close to each other but unlikely to have comparable traffic counts.
- Additional covariates such as number of lanes are important to distinguishing subtypes of A4 roads.
- There might also be measurement errors or other data quality issues for A4 roads. The last example illustrates an unusual case.

A4 Map View: Pair I [13]

- Red: Woodland Square Loop SE (traffic count: 1746)
- Blue: College St SE (traffic count: 50436)

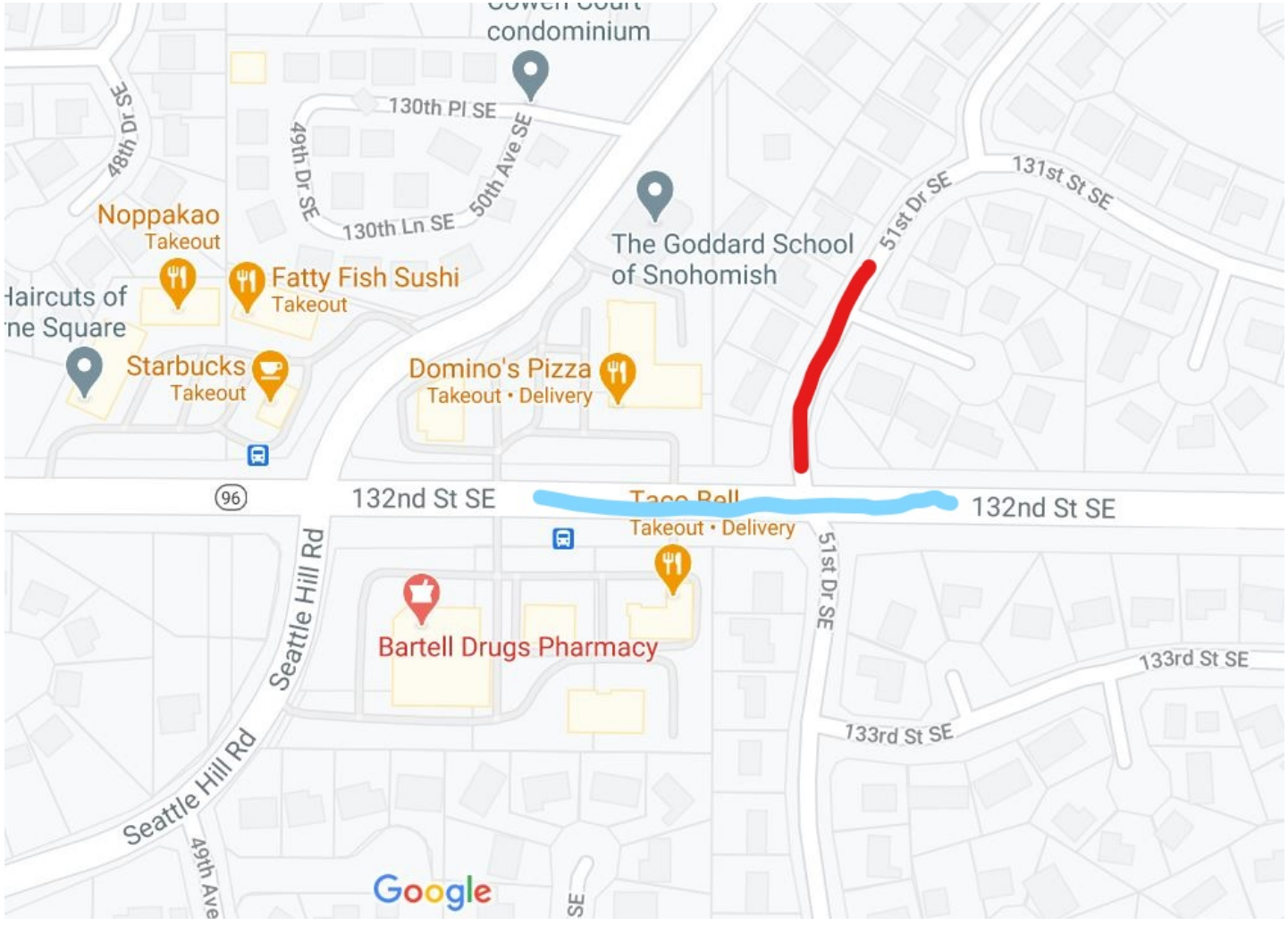


[13] all pictures are screen shots from Google Map

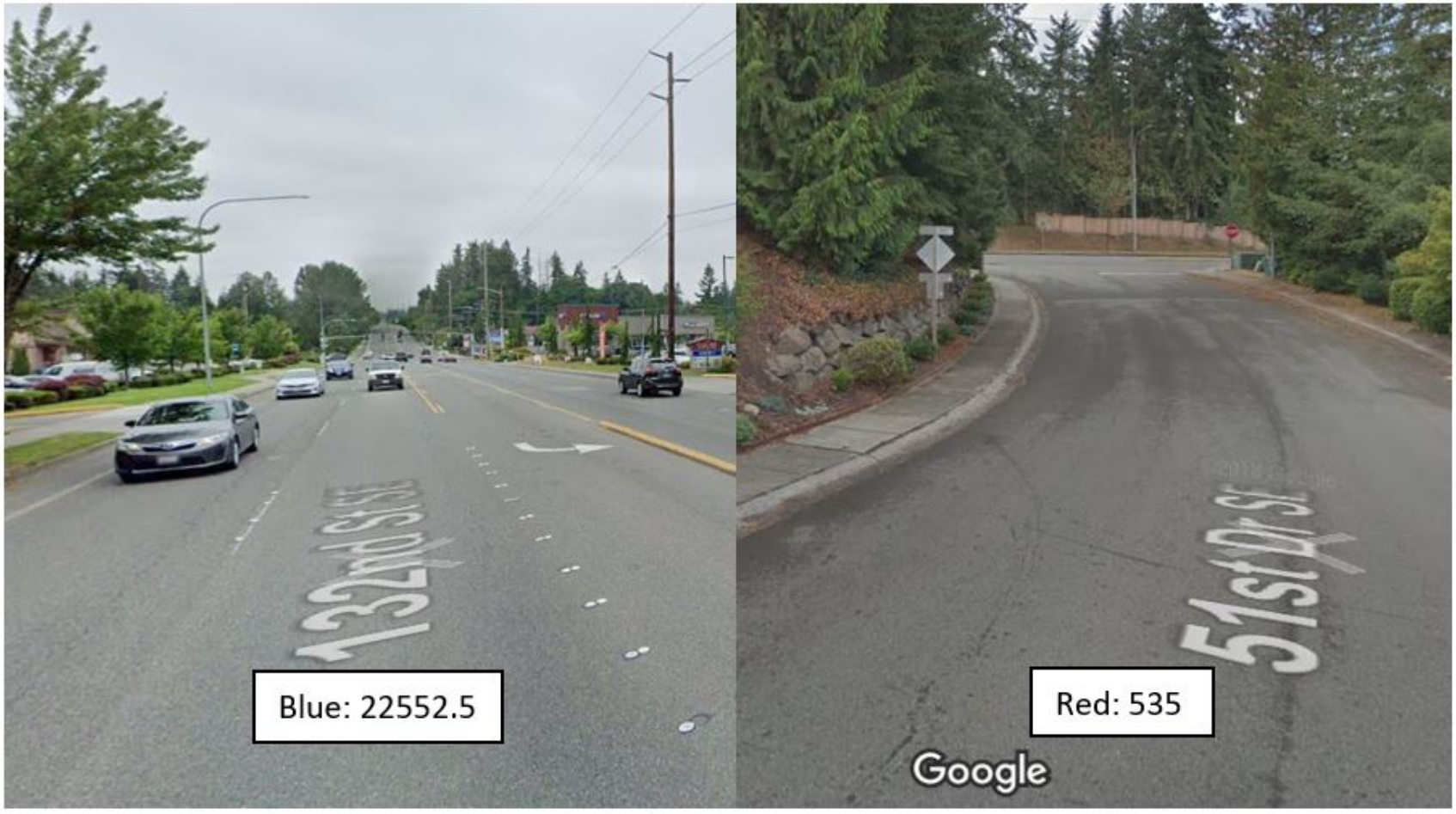
A4 Street View: Pair I

A4 Map View: Pair II

- Red: 51st Dr SE (traffic count: 535)
- Blue: 132nd St SE (traffic count: 22552.5)

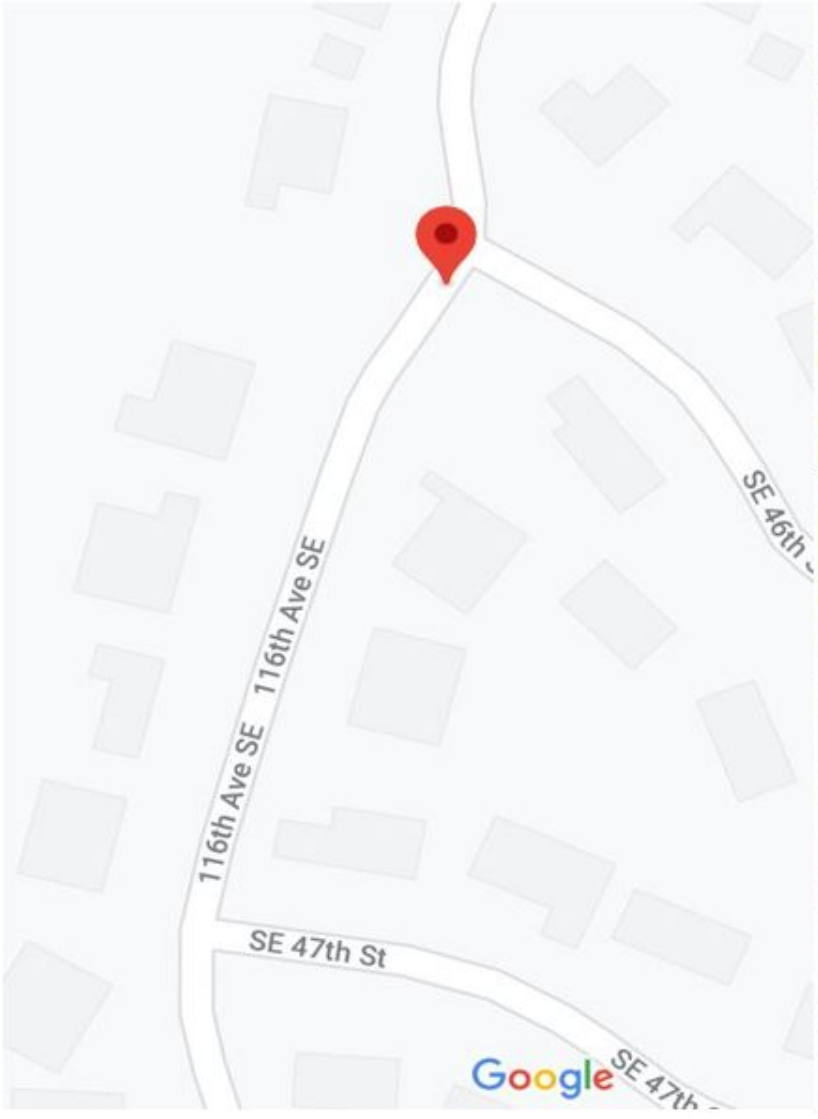


A4 Street View: Pair II



A4 Unusual Example

Traffic count on 116th Ave SE: 149549.



Next Steps

1. Improve A4 performance by adding new covariates and modifying model.
2. Develop a metric to assign traffic count at roads without observations.
3. Apply the method to all roads across US.

Reference

- Finley, A.O., Datta, A., Banerjee S. (2020) spNNGP R package for Nearest Neighbor Gaussian Process models. accessed from https://arxiv.org/pdf/2001.09111.pdf

- Finley, A.O., Datta, A., Cook, B.C. *et al*. (2019) Efficient algorithms for Bayesian nearest-neighbor Gaussian processes. *Journal of Computational and Graphical Statistics*. **28:2**, 401-414, doi: 10.1080/10618600.2018.1537924

- Heaton, M.J., Datta, A., Finley, A.O. *et al*. (2019) A Case Study Competition Among Methods for Analyzing Large Spatial Data. *JABES* **24**, 398–425. https://doi.org/10.1007/s13253-018-00348-w

- Wood SN (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, **65(1)**, 95-114.