Multilevel time series modeling of mobility trends in the Netherlands for small domains

Sumonkanti Das[1,3], Harm Jan Boonstra[2], and Jan van den Brakel[1,2]

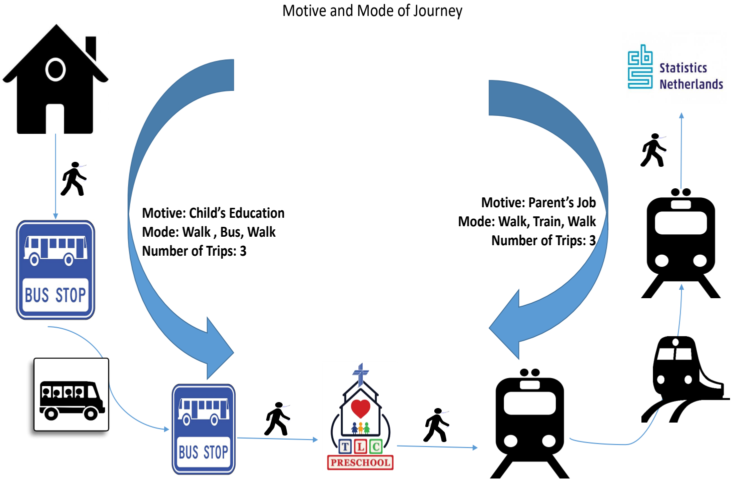
[1] Maastricht University, The Netherlands
[2] Statistics Netherlands, The Netherlands
[3] University of Wollongong, Australia

ITACOSM 2019 - Survey and Data Science
University of Florence
June 5, 2019 June 7, 2019

Introduction

- Main purpose of the Dutch Travel Survey (DTS) is to produce reliable estimates on mobility of the Dutch population.
- Here three mobility characteristics per person per day (pppd) are considered
  - Average number of journey legs pppd (anjl-pppd)
  - Average distance per journey leg (adjl)
  - Average distance pppd (ad-pppd) based on anjl-pppd and adjl
- Journey legs are characterized by journey motive and transportation modes for a particular journey.

Journey legs by motive and transportation modes



Aims of the study

- At first, trends of mobility indicators for the period 1999-2017 are aimed to estimate for small domains
- $D = 504$ small domains are cross-classification of:
    - sex (male, female)
    - ageclass (0-5, 6-11, 12-17, 18-29, 30-39, 40-49, 50-59, 60-69, 70+)
    - motive (work, shopping, education, other)
    - mode (car driver, car passenger, train, BTM (bus/tram/metro), cycling, walking, other)
- Finally, predictions of trends for higher aggregation levels by aggregation of most detailed level predictions
    - Aggregation Level (say): Overall, Motive, Mode, Motive X Mode

Predictions of Mobility Trends

- Time series multilevel models (TSMMs) are defined at the most detailed level (i.e., sex-ageclass-motive-mode)
- TSMMs for small area prediction are extensions of the area level Fay-Herriot (1979) model.
- Direct estimates of anjl-pppd and adjl and their estimated standard errors (SE) are utilized as input
- TSMMs are expressed in a hierarchical Bayesian framework and fit using a MCMC simulation method.

Problems in Predictions of Mobility Trends

- Discontinuities due to the redesigns of DTS in 2004 (from OVG to MON), and 2010 (from MON to OViN)
- These discontinuities are more visible at aggregate level and need to be accounted in modeling
- Small sample size to obtain reliable point estimates and stable SEs for many domains
- Outliers due to less reliable point estimates in 2009
- Many domains (without structurally zero domains) with zero direct estimates due to no observations of trip legs.

Some Notations

- $\hat{Y}_{it}$ = Direct estimate for year t and domain i
- $se(\hat{Y}_{it})$ = Estimated standard error of $\hat{Y}_{it}$
- Generalized Variance Function (GVF) model is developed for getting smoothed $se(\hat{Y}_{it})$
- $\hat{\theta}_{it}$ = Trend estimates resulting from developed TSMMs
- $se(\hat{\theta}_{it})$ = Estimated SE of $\hat{\theta}_{it}$ resulting from the TSMMs
- $\hat{\theta}_{it}$ and $se(\hat{\theta}_{it})$ are MCMC approximations of the posterior mean and standard deviation

Multilevel time series model

The initial estimates $\hat{Y}_{it}$ are combined into a M-vector as

$$\hat{Y} = (\hat{Y}_{11}, \dots \hat{Y}_{M_d 1}, \dots \hat{Y}_{1T}, \dots \hat{Y}_{M_d T})$$

where $M_d = 504$, $T = 19$ and $M = M_d \times T$. The multilevel models take the general linear additive form

$$\hat{Y} = X\beta + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha)} + e$$

- X = $M \times p$ design matrix for a p-vector of fixed effects β
- $Z^{(\alpha)}$= $M \times q^{(\alpha)}$ design matrices for $q^{(\alpha)}$-dimensional random effect vectors $v^{(\alpha)}$
- Sampling errors $e = (e_{11}, \dots, e_{M_d 1}, \dots e_{M_d T}) \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \Phi = \oplus_{t=1}^{T} \Phi_t$ with Φ_t= $COV(\hat{Y}_{1t}, \dots, \hat{Y}_{M_d t})$. Here, Φ_t is assumed diagonal.

Multilevel time series model

- $v^{(\alpha)}$ for different α are assumed to be independent
- However, components of a vector $v^{(\alpha)}$ can be correlated to accommodate temporal / cross-sectional correlation.
- For convenience, the superscript α is suppressed later.
- Each vector v is assumed to be distributed as

$$v \sim \mathcal{N}(0, A \otimes V)\,,$$

  where V and A are $d \times d$ and $l \times l$ covariance matrices.
- Length of v is $q = dl$
  - Simply d effects allowed to vary over l levels of a factor.
  - e.g. motive ($d = 4$) varying over time ($l = 19$ years).

Multilevel time series model

- Matrix V is allowed to be parameterized as
  - fully parameterized (unstructured) covariance matrix,
  - a diagonal matrix with different elements (diagonal)
  - a diagonal matrix with equal elements (scalar).
- Modelling multiple varying effects: Choose an appropriate covariance matrix V
- Generalisation of V to non-normal distributions of random effects
  - Student-t, Horseshoe prior, Laplace distributions

Multilevel time series model

- Matrix A describes known covariance structure between the levels of a factor variable.
- Precision matrixces $Q_A = A^{-1}$ instead of A are used.
- Modelling variations over time: Choose a Matrix A with appropriate correlation structure
  - First-order random walk (RW1): Local level trends
  - Second-order random walk (RW2): Smooth trends
- Generalisation of A to non-normal distributions can be done as V

Multilevel time series model

- Models are fitted using MCMC sampling (Gibbs sampler)
- Specification of the full conditional distributions are available in Boonstra and Brakel (2018).
- Model selection procedure: Widely Applicable Information Criterion (WAIC) and Deviance Information Criterion (DIC).
- The models are run in R using package `mcmcsae`.
- A longer run of 1000 burn-in plus 10000 iterations. Finally, 3 chains $\times$ 2000 iterations $=$ 6000 draws to compute estimates and standard errors.

Model Development: Average number of journey legs pppd (anjl-pppd)

- SQRT transformation of $\hat{Y}_{it}$ and $se(\hat{Y}_{it})$
- Taylor approximation of $se(\hat{Y}_{it}) \rightarrow se(\hat{Y}_{it})/(2\sqrt{\hat{Y}_{it}})$
- GVF model is applied to the transformed $se(\hat{Y}_{it})$
- Covariates along with `sex`, `ageclass`, `motive`, `mode`
  - `br_mon`: takes values 1 for 2004-2009 years
  - `br_ovin`: takes values 1 for 2010-2017 years
  - `dummy_2009`: Binary variable for year 2009
  - `yr.c`: Scaled and centered version of year `(yr)`
  - `br_mon_SO`: Equal to `br_mon` for motives shopping & other
  - `snowdays`: Annual number of snowdays
- Final time series model includes fixed and random effects

Model Development: anjl-pppd

- Fixed effects components:
  $sex * ageclass + motive * mode + (ageclass + motive + mode) * (br\_ovin + yr.c) + mode * snowdays$
- Random effects component:

Model Component	Formula V	Variance Structure	Factor A	PriorA	Number of Effects[1]
V_2009	$dummy\_2009$	scalar	$sex * ageclass * motive * mode$	Horseshoe	504
V_BR	$1 + yr.c + br\_mon\_SO + br\_ovin$	unstructured	$sex * ageclass * motive * mode$	Laplace	1764
RW2AMM	$ageclass * motive * mode$	scalar	RW2(yr)	normal	4788
RW2MM	$motive * mode$	diagonal	RW2(yr)	normal	532
RW1SAM	sex	unstructured	$ageclass * mode *$ RW1(yr)	normal	2394
WN	1	scalar	$sex * ageclass * motive * mode * yr$	normal	9576

- Normal distribution for the sampling errors works better

[1] This includes effects for structural zero domains.

Model Development: Average distance per journey leg (adjl)

- LOG transformation: $\hat{Y}_{it} \rightarrow \log(\hat{Y}_{it})$
- Taylor approximation of $se(\hat{Y}_{it}) \rightarrow se(\hat{Y}_{it})/(\hat{Y}_{it})$
- GVF model is applied to the transformed $se(\hat{Y}_{it})$
- Extra covariates
  - `log_ratio_km_NAP`: Logarithm of year-by-year differences of Car Kilometers registered in National Autopas (NAP)

Model Development: adjl

- Fixed effects component: $sex + ageclass + motive * mode + yr.c * mode + (mode\_walking + mode\_other) : br\_ovin + mode\_cardriver : log\_ratio\_km\_NAP$
- Random effects component:

Model Component	Formula **V**	Variance Structure	Factor **A**	PriorA	Number of Effects
V_BR	$1 + yr.c + br\_mon + br\_ovin$	unstructured	$sex * ageclass * motive * mode$	Laplace	2016
RW2M	*mode*	diagonal	RW2(yr)	normal	532
WN	1	scalar	$sex * ageclass * motive * mode * yr$	normal	9576

- Student-t distribution with $df = 4$ for the sampling errors works better

Results: Model Predictions and Trend Estimates

- Model predictions: Based on the linear predictor containing all model components

$$\eta^{(r)} = X\beta^{(r)} + \sum_{\alpha} Z^{(\alpha)} v^{(\alpha,r)} ,$$

where superscript r indexes the retained MCMC draws.

- Trend estimates of main interest: The level break effects and the dummy effects for outliers (if present) are removed from $\eta^{(r)}$

Average number of journey legs pppd: Overall Level

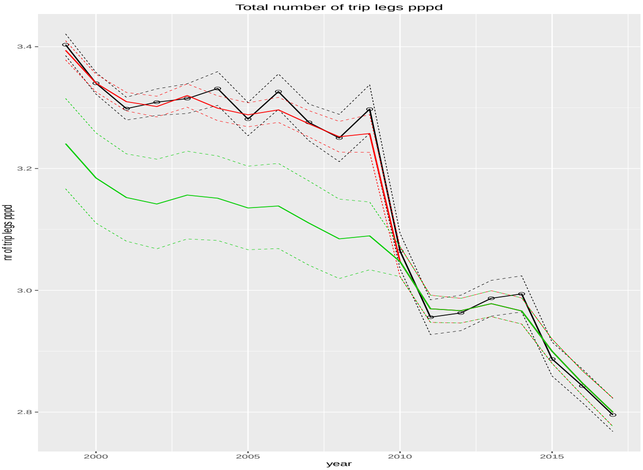


Figure: Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

Average number of journey legs pppd: Motive Level

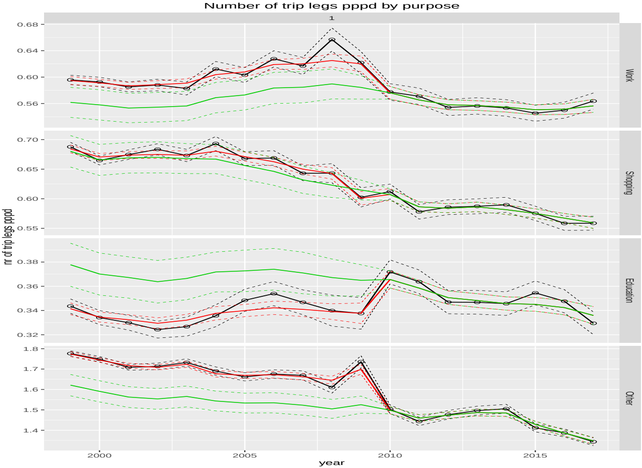


Figure: Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

Average distance per journey leg: Overall Level

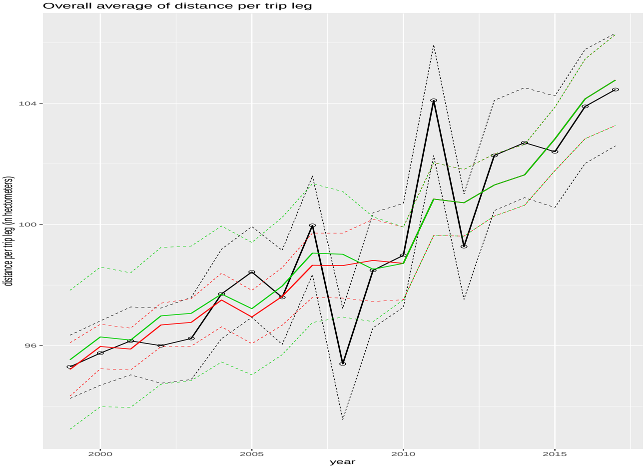


Figure: Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

Average distance per journey leg: Motive Level

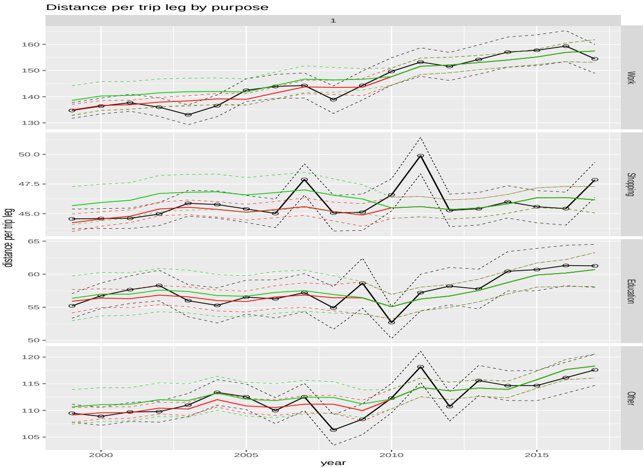


Figure: Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

Average distance pppd: Overall Level

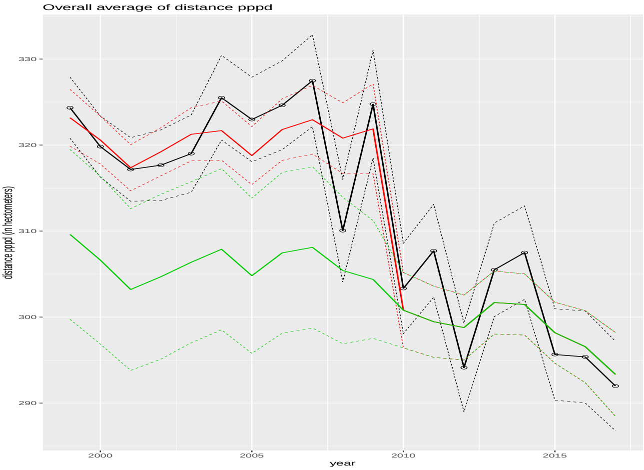


Figure: Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

Average distance pppd: Motive Level

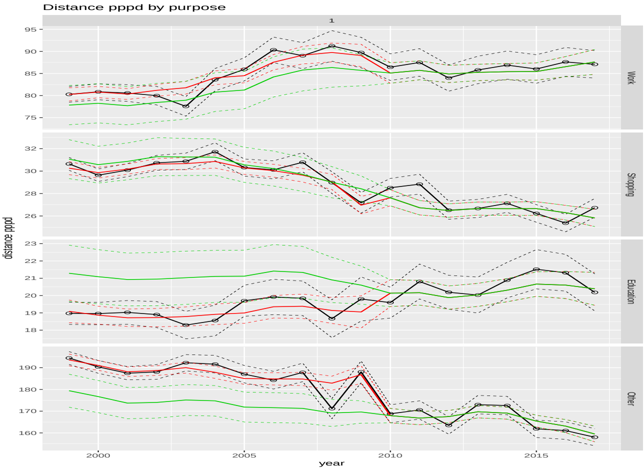


Figure: Direct estimates (black), model fit (red) and trend estimates (green) with approximate 95% intervals.

Model Diagnostics

- Normality and heteroskedasticity assumptions are checked at overall and detailed level.
- Autocorrelation of residuals is checked at detailed level.
- Posterior predictive check by calculating the posterior predictice p-values (PPP) for various statistics including weighted mean and variance.
- Model diagnostics confirm validity of the fitted multilevel time-series models for anjl-pppd and adjl

Concluding Remarks

- The developed time series models for anjl-pppd and adjl at the most-detailed level provide consistent trend estimates at different aggregation levels
- The models for anjl-pppd and adjl also provide consistent trend estimates of ad-pppd.
- The final models cover the possible critical issues of unstable SEs, effect of redesigns at several aggregation levels, and outliers

Concluding Remarks

- Outliers in the input estimates are accounted by considering t-distribution for sampling errors.
- Global-local shrinkage allows for some large random effects, while shrinking most (the noisy ones) to zero.
- Higher aggregation level variations are also accounted by incorporating some contextual variables.
- The similar model development procedure can be easily replicated to incorporate the new data of upcoming DTS (ODIN), which are based on new sampling design.

Reference

- Boonstra, H. J. and J. van den Brakel (2018). Hierarchical bayesian time series multilevel models for consistent small area estimates at different frequencies and regional levels. Statistics Netherlands discussion paper, December 4, 2018.
- Fay, R. and R. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74 (366), 269-277.
- Watanabe, S. (2013). A widely applicable bayesian information criterion. Journal of Machine Learning Research 14, 867-897.
- Wolter, K. (2007). Introduction to Variance Estimation. Springer.

Thank you for your patience

Appendix Table 1: Full covariance matrix for the component "V_BR"

- Est. standard deviations & correlations ($\times$100): anjl-pppd

	Intercept	br_mon_SO	br_ovin	yr.c
Intercept	14.76 (0.73)	3.97 (10.44)	-49.97 (5.55)	-7.52 (7.27)
br_mon_SO		1.74 (0.17)	24.88 (11.86)	28.37 (12.42)
br_ovin			3.25 (0.22)	-13.37 (8.40)
yr.c				1.38 (0.09)

- Est. standard deviations & correlations ($\times$100): adjl

	Intercept	br_mon	br_ovin	yr.c
Intercept	26.5 (1.4)	3.0 (27.8)	-19.6 (10.9)	17.7 (12.6)
br_mon		1.4 (0.9)	-2.6 (35.6)	17.1 (34.7)
br_ovin			11.8 (1.5)	0.6 (21.2)
yr.c				3.8 (0.7)

Appendix Table 2: Global and Local Scale Parameters for "V_2009" and "V_BR"

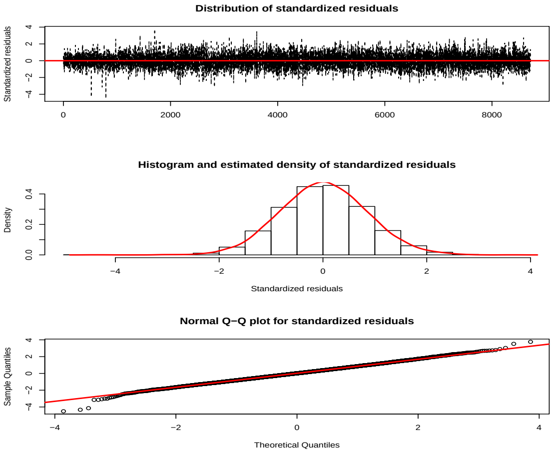
- σ_{V2009}=0.0036 with SE= 0.0009
- Summary statistics of the local scale parameters: anjl-pppd

Component	Min	$Q_{0.25}$	Median	Mean	$Q_{0.75}$	Max
V_2009	2.69	4.46	5.52	34.0	9.19	4710
V_BR	0.44	0.65	0.84	0.91	1.05	2.16

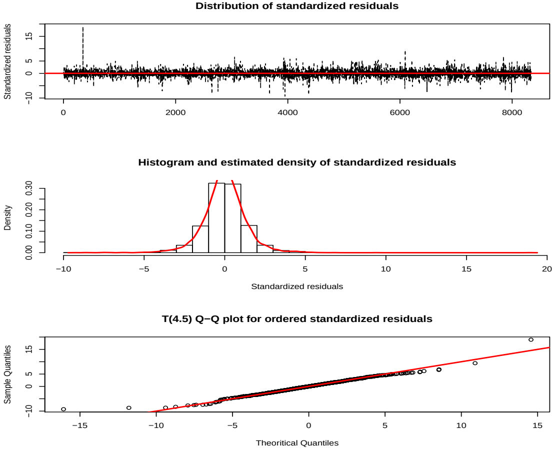
- Summary statistics of the local scale parameters: adjl

Component	Min	$Q_{0.25}$	Median	Mean	$Q_{0.75}$	Max
V_BR	0.52	0.77	0.94	0.96	1.08	2.24

Appendix Figure 1: Residual diagnostics of the standardized residuals for anjl-pppd



Appendix Figure 2: Residual diagnostics of the standardized residuals for adjl



Appendix Table 3: Normality, homoskedasticity and serial correlations of domain-specific residuals

Proportion of domains for which the standardized residuals satisfy the assumptions of normality, student-t (df=4.5), homoskedasticity and serial correlations

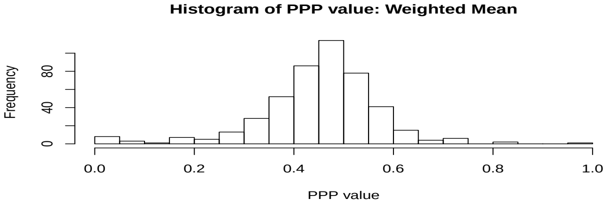
- For anjl-pppd

	Normal	Homoskedasticity	Serial Correlation	Total Domain
Proportion	97.37	92.11	12.72	456

- For adjl

	Normal	Student-t	Homoskedasticity	Serial Correlation	Total Domain
Proportion	89.32	99.55	88.41	9.09	440.00

Appendix Figure 3: PPP for weighted mean and variance for anjl-pppd



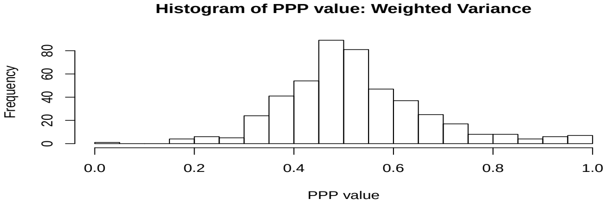


Figure: Distribution of PPP at detailed level

Appendix Figure 4: PPP for weighted mean and variance for adjl

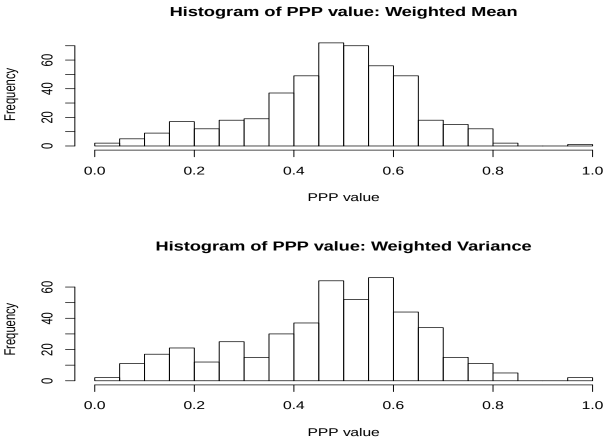
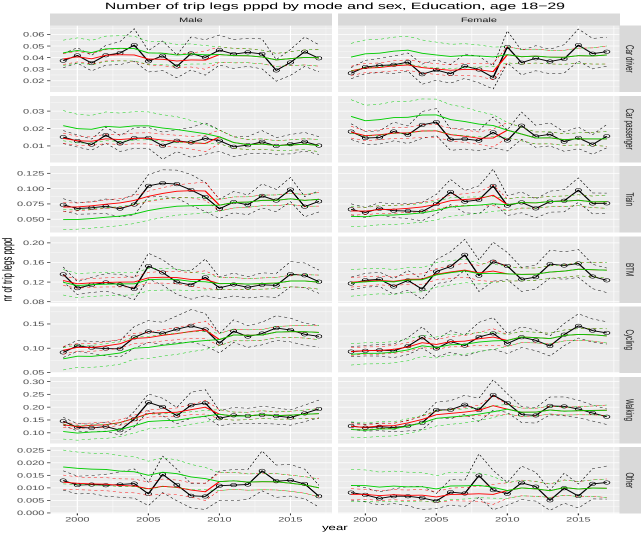
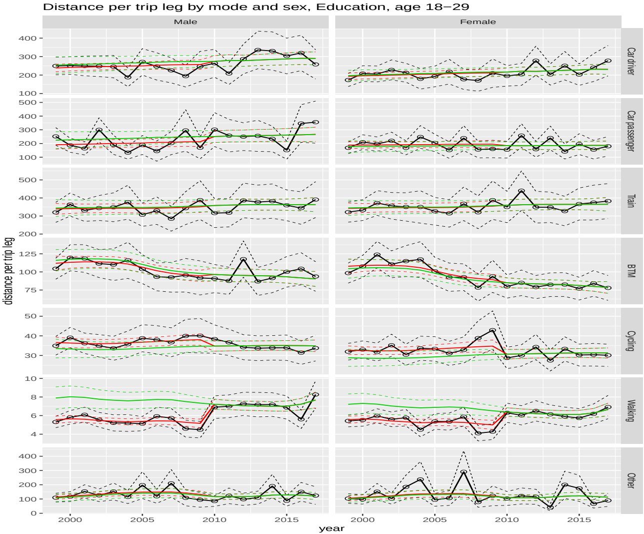


Figure: Distribution of PPP at detailed level

Appendix Figure 5: Prediction at detailed level for anjl-pppd



Appendix Figure 6: Prediction at detailed level for adjl



Appendix Figure 7: Prediction at detailed level for ad-pppd