Multilevel Models and Missing Data Models for Crowdsourced Bicycle Route Ratings

Will Jones

May 2, 2016

The Data

Ride Report

Ride collection:

- route, timestamp, length collected automatically
- ratings provided by user

Three main issues in data (we are concerned with here):

1. Subjective ratings
2. Many ratings missing
3. Many rides misclassified

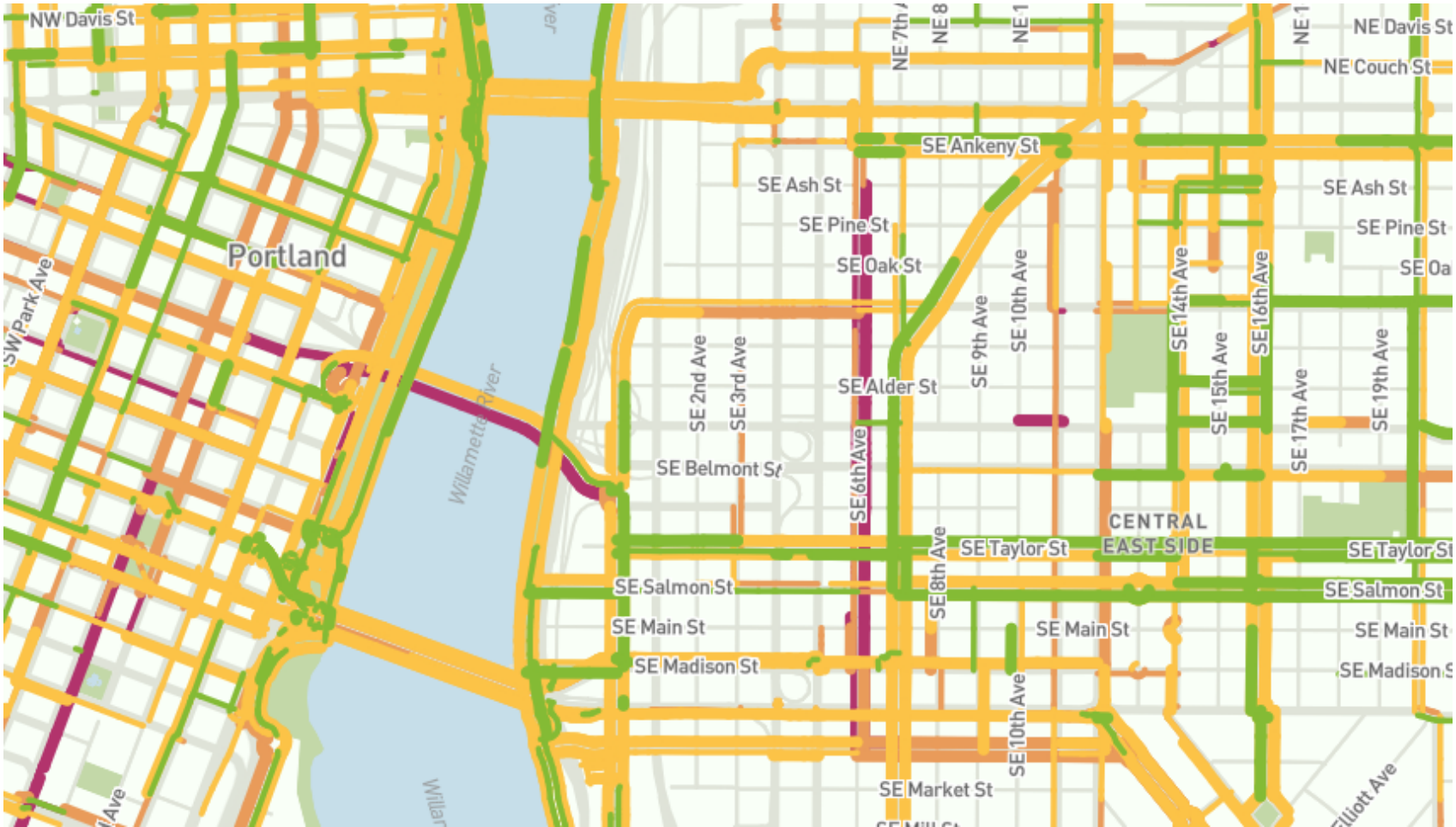
Other issues:

- Some rides are split
- No obvious way to model routes

Why study this data?

Ride Report's Goals

- **For cities:** identify the problematic street segments in the city for urban cyclists
- **For cyclists:** identify the best routes for commuting by bike



Our Approach

- We attempt to address:

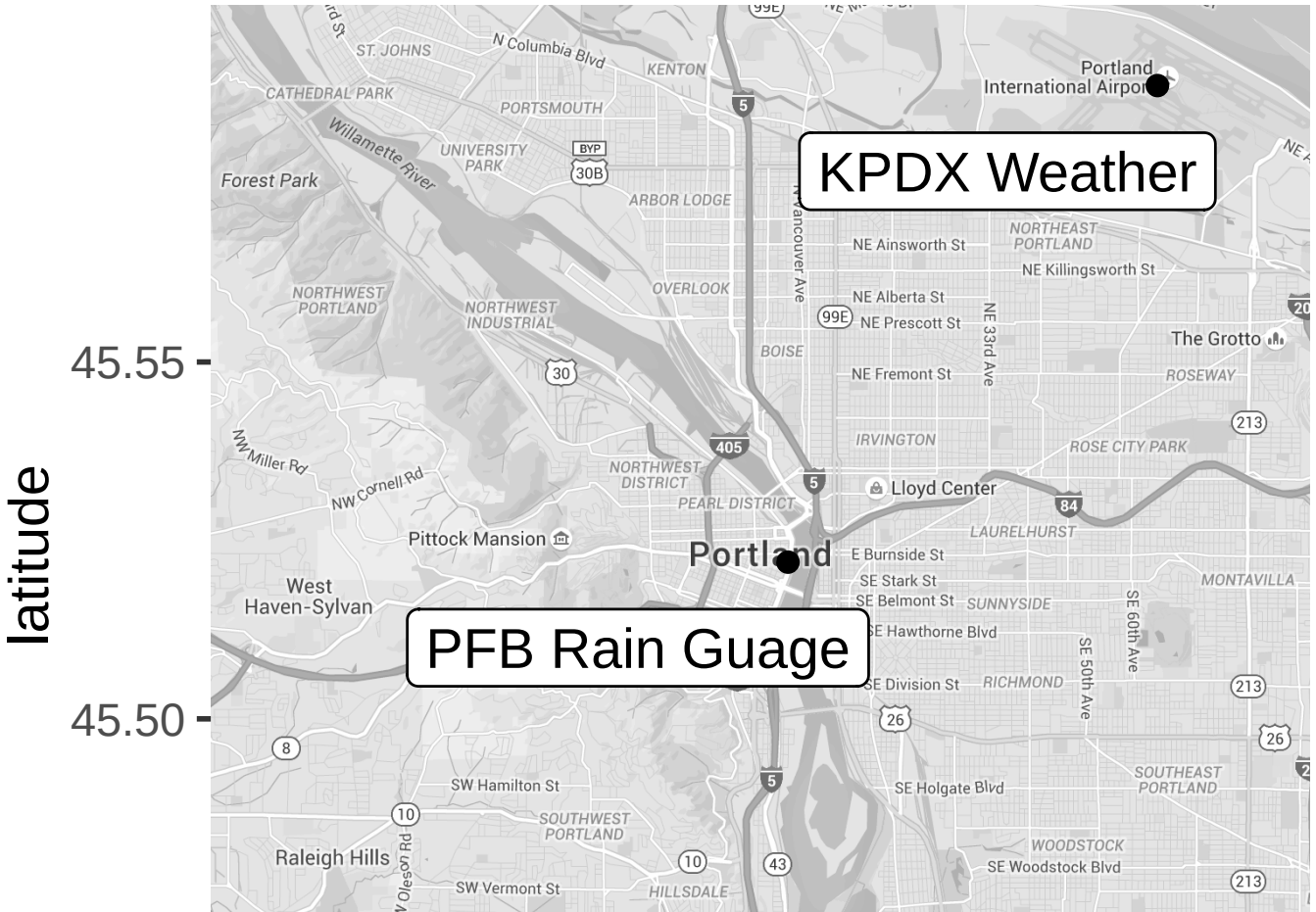
1. Issue 1 (subjective ratings) with multilevel models
2. Issue 2 (missing ratings) with the expectation maximization algorithm

- We don't use route information; further research needed

Weather Data Sources

We combined the ride data with

- daily weather data from the KPDX weather station (Weather Underground, 2016)
- hourly rain guage data from the Portland Fire Bureau rain guage (Portland Bureau of Environmental Science, 2016)



A note about reproducibility

This entire analysis is availible as an R package in a GitHub repository

- URL: `https://github.com/wjones127/thesis`
- But, Ride Report data is not included

Ride Rating Models

Notation

We have n observations of rides.

$$y_i = \begin{cases} 1, & \text{if ride } i \text{ was given a negative rating;} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Define predictors,

- x_i^{length}, log length of ride
- x_i^{rain}, rainfall during hour of ride
- x_i^{rain4h}, cumulative rainfall from past four hours
- x_i^{wind}, mean wind speed that day
- x_i^{gust}, maximum gust speed
- x_i^{temp}, mean temperature that day
- t_i, time of day of ride

for $i = 1, \ldots, n$. All but the last we represent together as matrix X.

Six Models for Ride Rating

- **Model 1**: Logistic regression model
- **Model 2**: Add rider random intercepts
- **Model 3**: Add trigonometric terms for t
- **Model 4**: Additive multilevel model with cyclic cubic spline for t
- **Model 5**: Model 4 with cubic splines for x^{length}
- **Model 6**: Model 4 with fixed intercept

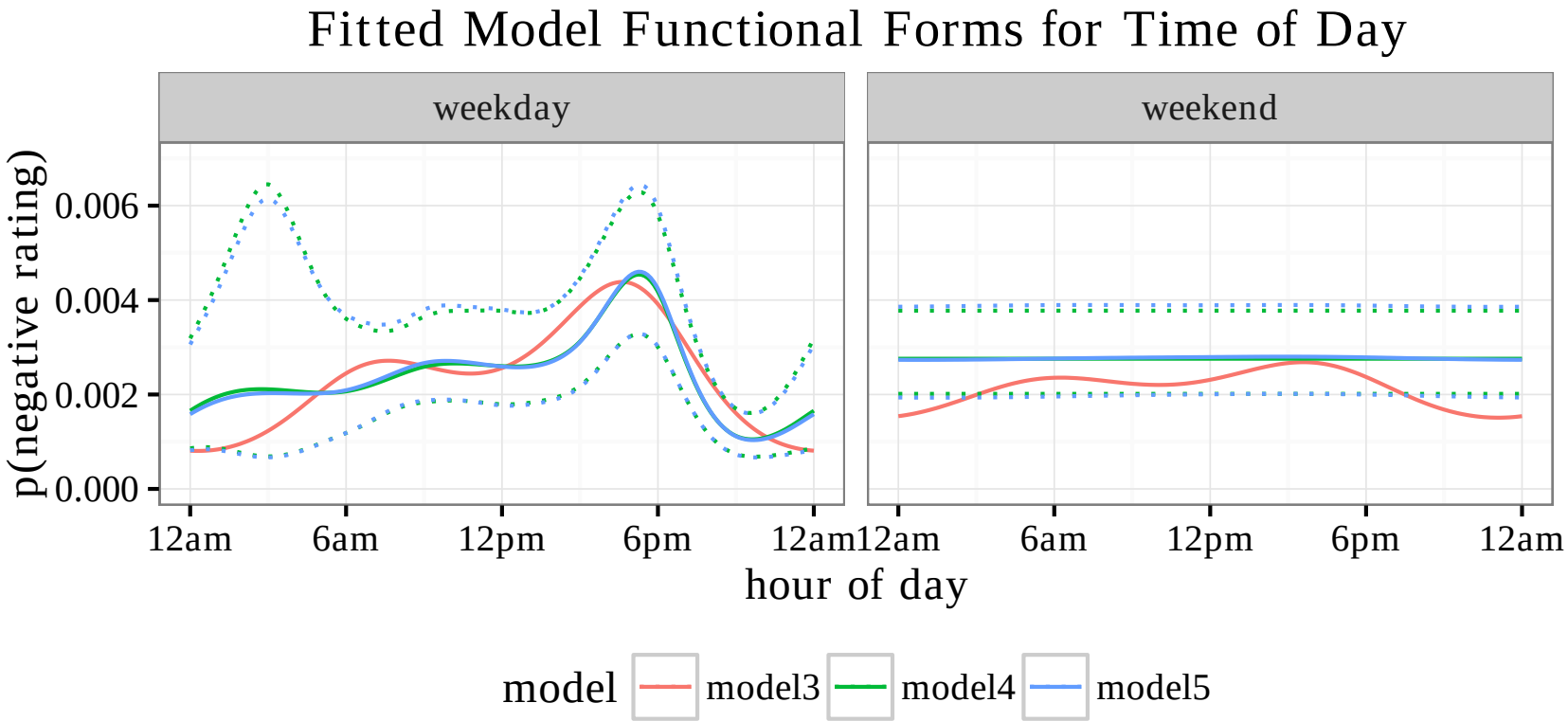
How well do they fit?

Table 1: Fit summaries for Models 1–6.

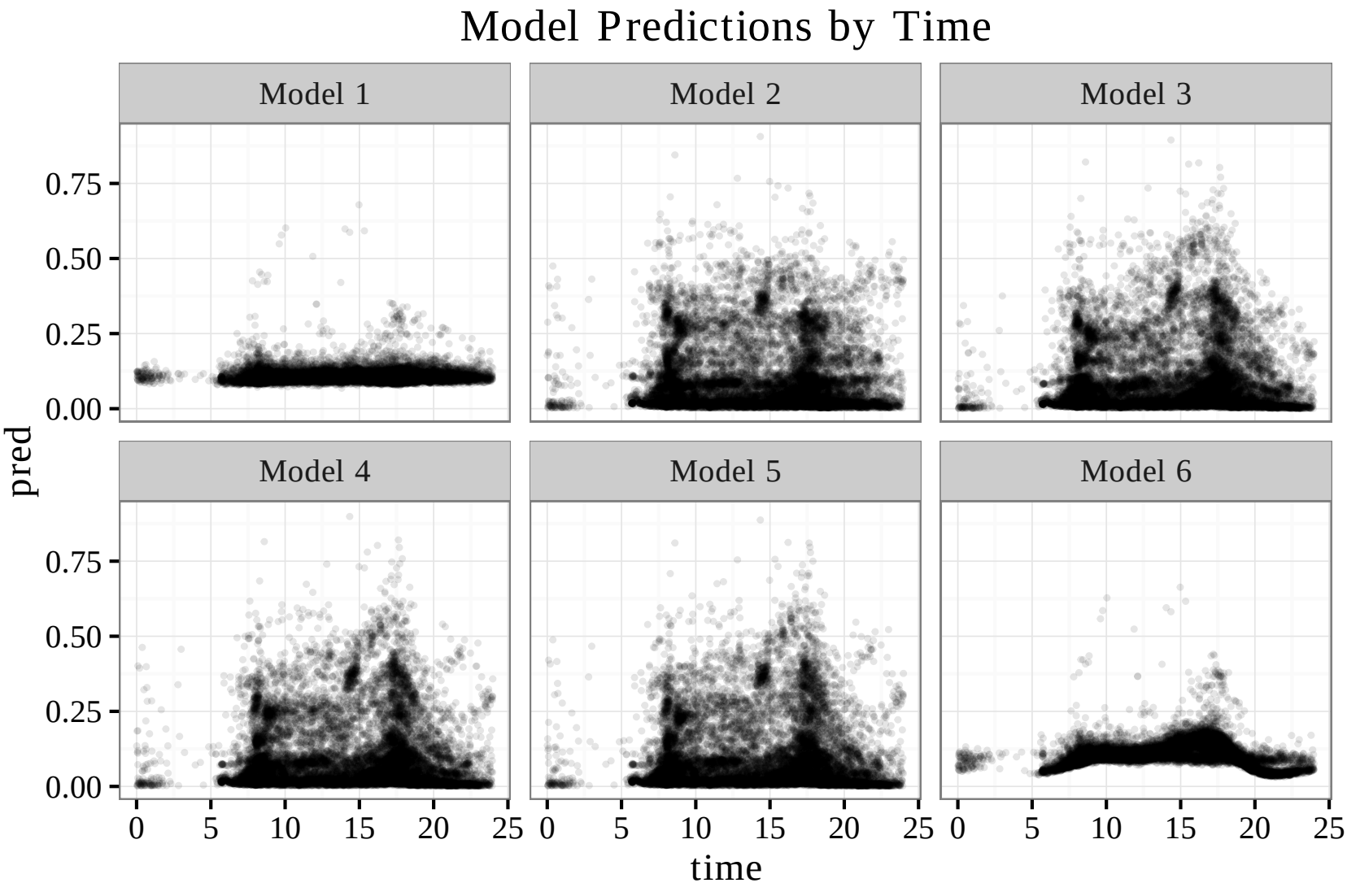
Model	$\log(\mathcal{L})$	**AIC**	**AUC**$_{\text{CV}}$[1]
Model 1	-4,786	9,586	0.552
Model 2	-3,971	7,957	0.797
Model 3	-3,923	7,877	0.802
Model 4	-3,930	7,870	0.802
Model 5	-3,928	7,878	0.803
Model 6	-4,713	9,455	0.601

[1] Area under ROC curve estimated with 10-fold cross-validation.

Time of Day Trends



What do the intercepts encode?



Classifying Riders

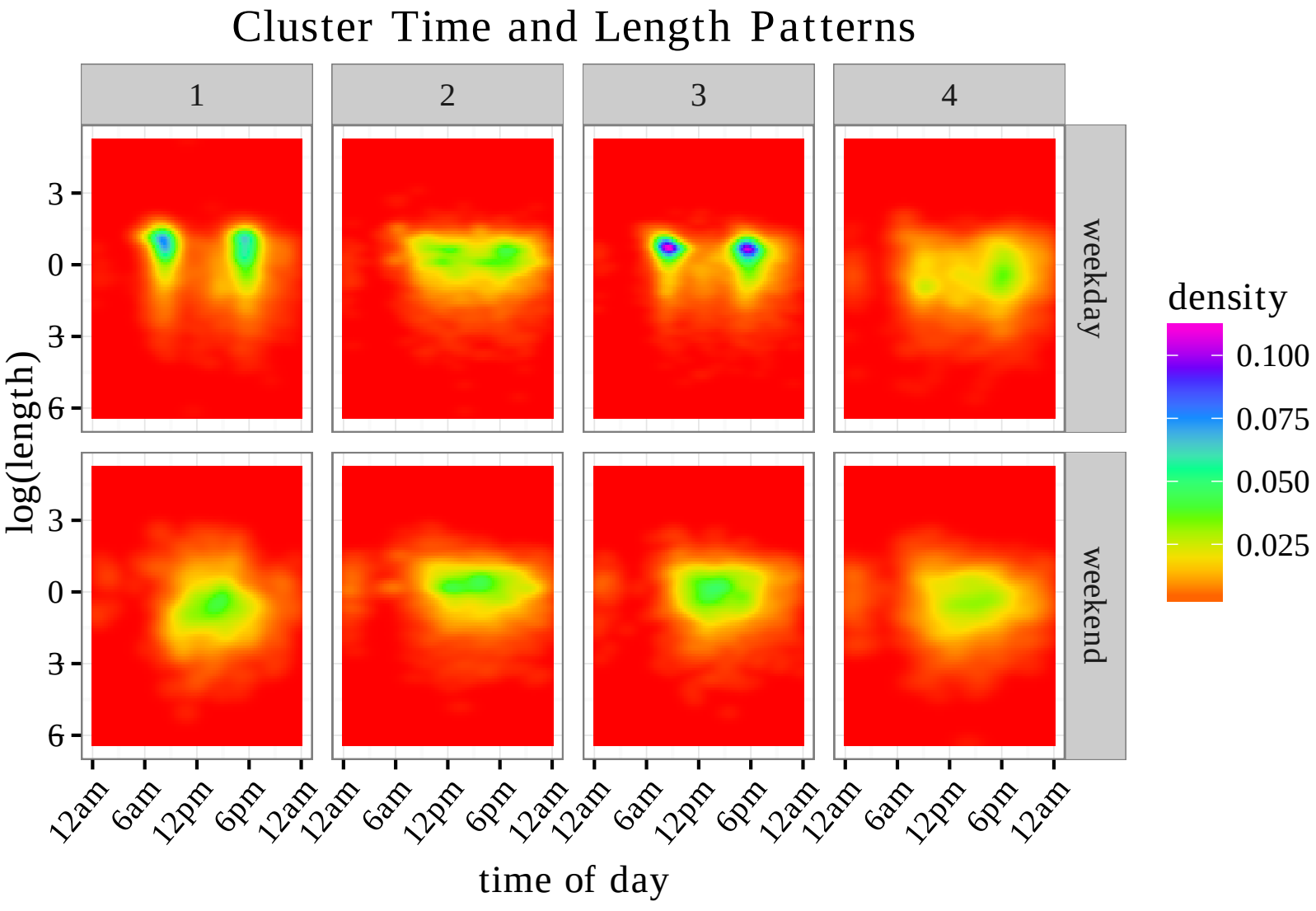
What features can we use?

For riders $j = 1, \ldots, I$, we have rider-level predictors

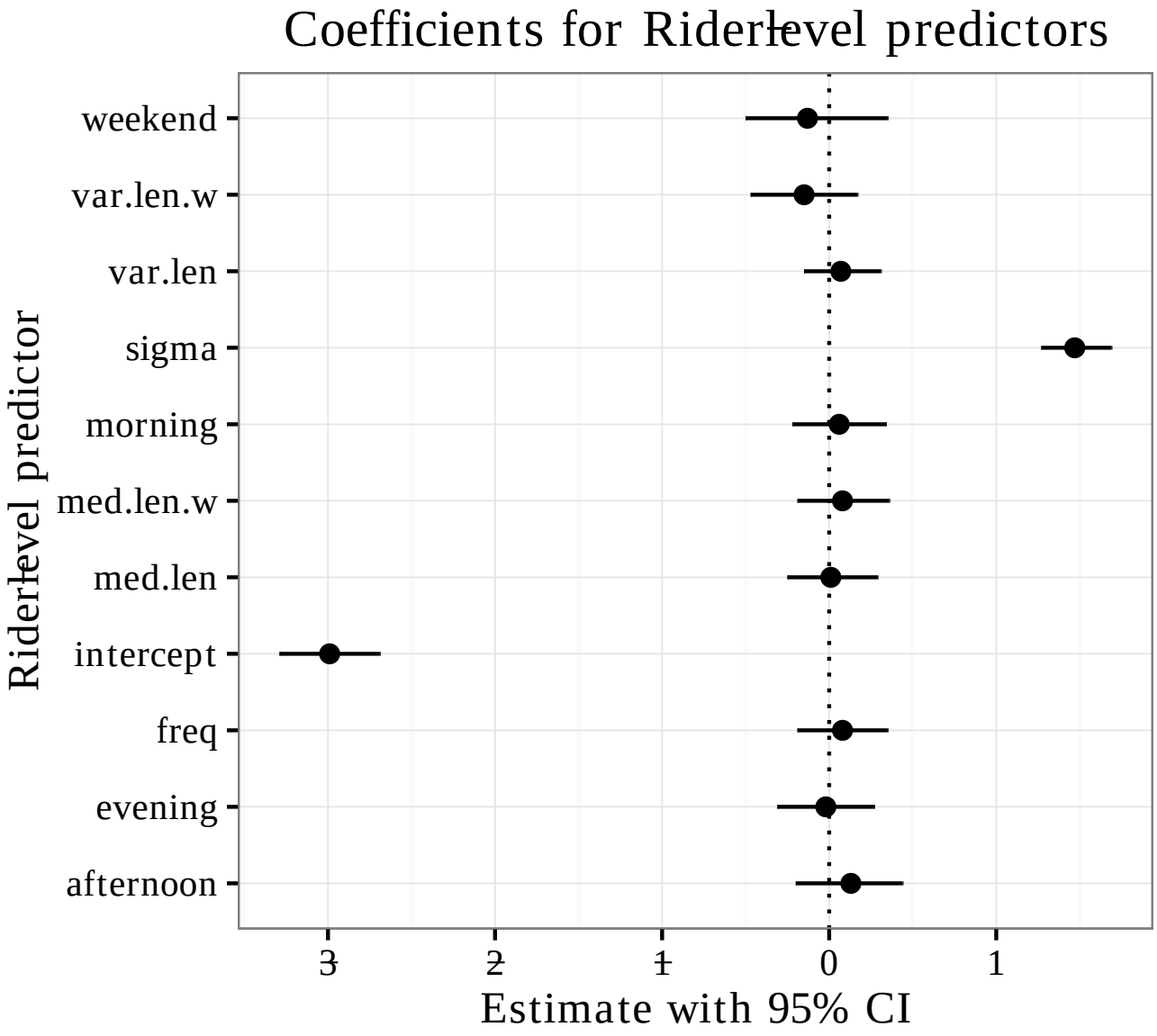
- u_j^{freq}, frequency of rides
- u_j^{weekend}, proportion of rides on weekends
- $u_j^{\text{med.len}}$, median length of weekday rides
- $u_j^{\text{med.len.w}}$, median length of weekend rides
- $u_j^{\text{var.len}}$, variance of length of weekday rides
- $u_j^{\text{var.len.w}}$, variance of length of weekend rides
- u_j^{morning}, proportion of rides in morning
- u_j^{lunch}, proportion of rides during lunch
- u_j^{evening}, proportion of rides in evening

Variables were standardized and clustered using k-means clustering

What patterns do these riders exhibit?



What good are these as predictors for rider intercepts?



Rider clusters and rider-level predictors

- Model using rider-level predictors to predict intercept found they were poor predictors
- Model using cluster intercepts performed much worse than rider intercepts
- Unsupervised learning methods are an art; maybe there is a better way

Missing Data

Missing Ratings

Of $n = 25,397$ rides, $11,365$ **not rated**.

- Is it safe to ignore these observations?

Types of Missing Data

Let,

$$r_i = \begin{cases} 1, & \text{if ride } i \text{ is missing a rating;} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Rubin classifies missing data into three situations[2]:

1. **Missing Completely at Random (MCAR)**, where r is independent of r and the predictors X. *i.e.* $\mathbb{P}(r = 1|y, X) = \mathbb{P}(r = 1)$
2. **Missing at Random (MAR)**, where r is independent of y, but may depend on X, *i.e.* $\mathbb{P}(r = 1|y, X) = \mathbb{P}(r = 1|X)$
3. **Nonignorable, or not MCAR nor MAR**, where r is dependent on y.

We believe the missing ratings are nonignorable.

[2] Little & Rubin (1987) (page 14)

The EM Algorithm for Missing Data

- The Expectation Maximization (EM) algorithm: procedure for fitting models with latent variables.
- Here, latent variables are missing observations
- We use the weighting method proposed by Ibrahim and Lipsitz[3].

[3] Ibrahim & Lipsitz (1996)

The EM Algorithm: Setup

- Have **data model**: $f(y \mid X, \beta)$ and **missing data model**: $f(r \mid X, y, \alpha)$

EM algorithm general procedure[4]:

1. *E-step*: Compute expected loglikelihood,

$$Q(\alpha, \beta|\alpha^{(t)}, \beta^{(t)}) = \int l(\alpha, \beta|y) \cdot f(y_{\text{mis}}|\, y_{\text{obs}}, \alpha^{(t)}, \beta^{(t)})\, dy_{\text{mis}} \tag{3}$$

2. *M-step*: maximize $Q(\alpha, \beta|\alpha^{(t)}, \beta^{(t)})$ to get $(\alpha^{(t+1)}, \beta^{(t+1)})$

[4] Little & Rubin (1987)

EM Algorithm: Weighting procedure

1. Get initial estimates of α and β.
2. Compute weights

$$w_{i\,y_i}^{(t)} = \frac{f(y_i \mid x_i, \beta^{(t)}) f(r_i \mid x_i, y_i, \alpha^{(t)})}{\sum_{y_i \in \{0,1\}} f(y_i \mid x_i, \beta^{(t)}) f(r_i \mid x_i, y_i, \alpha^{(t)})}. \quad (4)$$

3. Create augmented data:
4. Fit data model and missing data model separately using augmented data
5. Repeat 2–4 until loglikelihood converges

EM Algorithm: Augmented Data

- Allows us to fit models with any packages that supports weighting observations

Figure 1: Creation of augmented data set for the weighted method of the EM algorithm for missing response data.

Original Data				**Augmented Data**			
y_i	x_i	r_i		y_i	x_i	r_i	w_i
1	2.4	0		1	2.4	0	1
0	1.3	0	$\rightarrow$	0	1.3	0	1
NA	-0.4	1		1	-0.4	1	0.2
				0	-0.4	1	0.8

Missing Data Model Results: Data Model

Parameter	**Model 4**	**EM Model**
Log(Length)	-0.147	0.205
	(-0.290, -0.005)	(0.106, 0.304)
Mean Temperature	0.142	0.100
	(0.004, 0.281)	(0.005, 0.196)
Mean Wind Speed	0.002	-0.026
	(-0.054, 0.057)	(-0.069, 0.016)
Max Gust Speed	-0.005	0.020
	(-0.031, 0.021)	(0.001, 0.039)
Rainfall	0.050	0.051
	(-0.017, 0.117)	(0.009, 0.093)
Rainfall 4-Hour	0.022	0.017
	(0.003, 0.041)	(0.003, 0.030)
Intercept	-2.792	-3.144
	(-3.334, -2.250)	(-3.604, -2.684)

Missing Data Model Results: Nonresponse Model

Parameter	Basic Model	EM Model
y	0.730	1.035
	(0.235, 1.224)	(0.493, 1.577)
Log(Length)	-0.297	-0.327
	(-0.362, -0.232)	(-0.393, -0.262)
Mean Temperature	0.200	0.139
	(0.139, 0.262)	(0.077, -0.262)
Mean Wind Speed	0.032	0.031
	(0.003, 0.060)	(0.001, 0.061)
Max Gust Speed	-0.003	-0.007
	(-0.016, 0.010)	(-0.021, 0.006)
Rainfall	0.007	-0.024
	(-0.028, 0.041)	(-0.057, 0.009)
Rainfall 4-Hour	-0.002	0.010
	(-0.012, 0.009)	(-0.001, 0.021)
Intercept	-0.927	-0.967
	(-1.124, -0.729)	(-1.163, -0.771)

Should we trust these results?

- Recall issue 4: some rides are misclassified as bike rides
- Perhaps most unrated rides are misclassified?
- $\Longrightarrow$ need to fix misclassification before missing data methods can be properly applied

Conclusions

What we've learned

- Allowing random intercepts for rider gives huge improvements in model performance
- Modeling missing data may be essential, but data quality of missing data must be fixed

What remains to be researched

- How do we create models that use routes?
- Can riders be clustered in a way that predicts their rider intercept?
- How will rider intercepts change when route is incorporated?

References I

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Ibrahim, J. G., & Lipsitz, S. R. (1996). Parameter estimation from incomplete data in binomial regression when the missing mechanism is nonignorable. *Biometrics*, *52*(3), 1071–1078.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Portland Bureau of Environmental Science, C. of. (2016, February). Portland fire bureau rain gage data table. Retrieved from

References II

http://or.water.usgs.gov/non-usgs/bes/ankeny.rain

Software, K. (2016). *Portland ride report rides: December 3, 2014 – april 8, 2016*.

Team, R. C. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Weather Underground, I. (2016, February). Weather history for kPDX. Retrieved from https://www.wunderground.com/history/airport/KPDX/2015/9/20/CustomHistory.html?dayend=8&monthend=2&yearend=2016&req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo=&MR=1

Wood, S., & Scheipl, F. (2014). Gamm4: Generalized additive mixed models using mgcv and lme4. Retrieved from https://CRAN.R-project.org/package=gamm4