Bus Travel Time Predictions using Additive Models

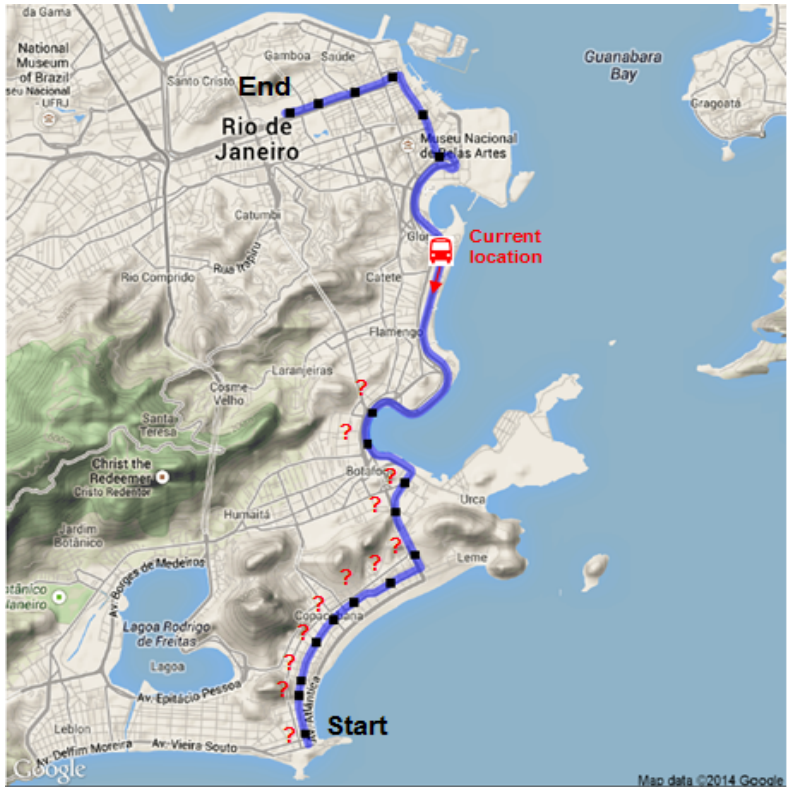
Matthías Kormáksson
IBM Research - Brazil

December 4, 2014

Overview

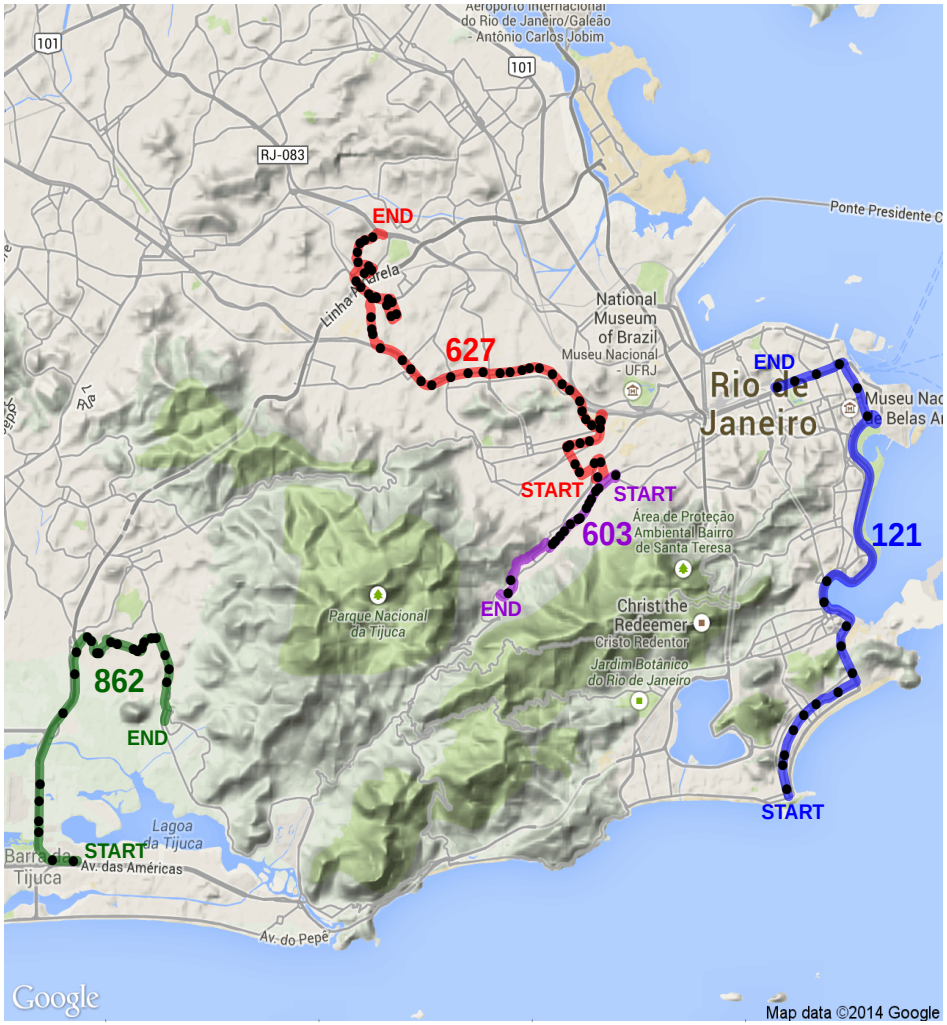
- Problem Statement
- Motivating Data
- Background
- Proposed Solution
- Experiments
- Conclusions

Problem Statement



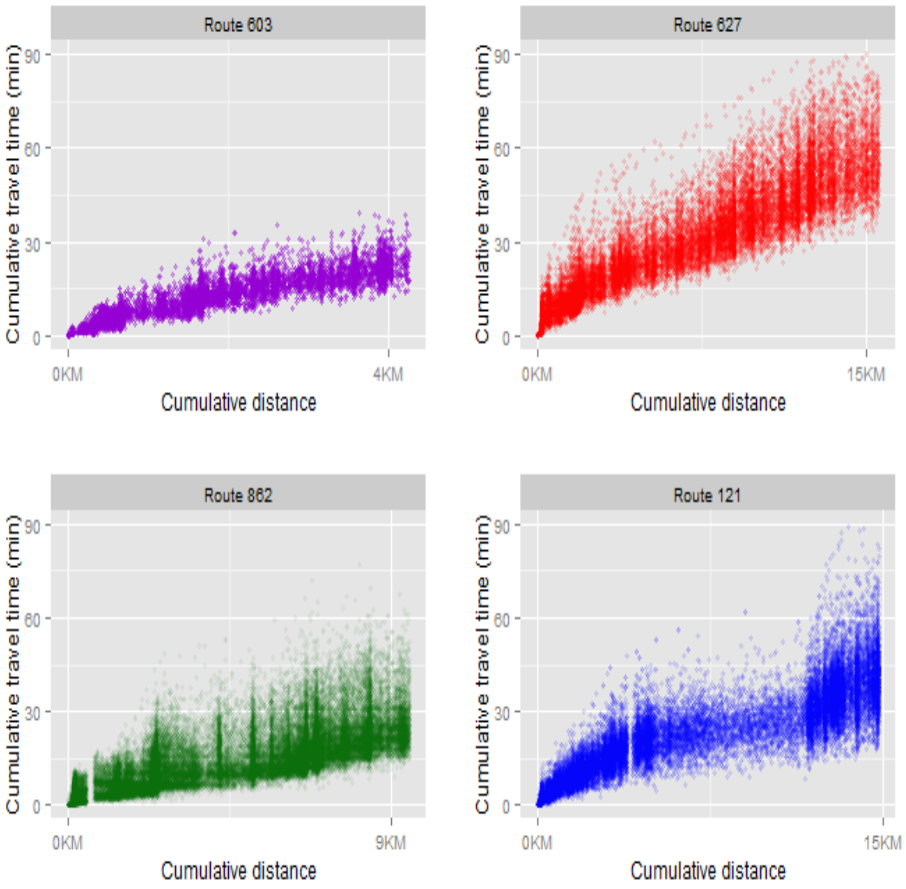
Given current time and location of the bus, what are the arrival times at subsequent bus stops?

Motivating Data



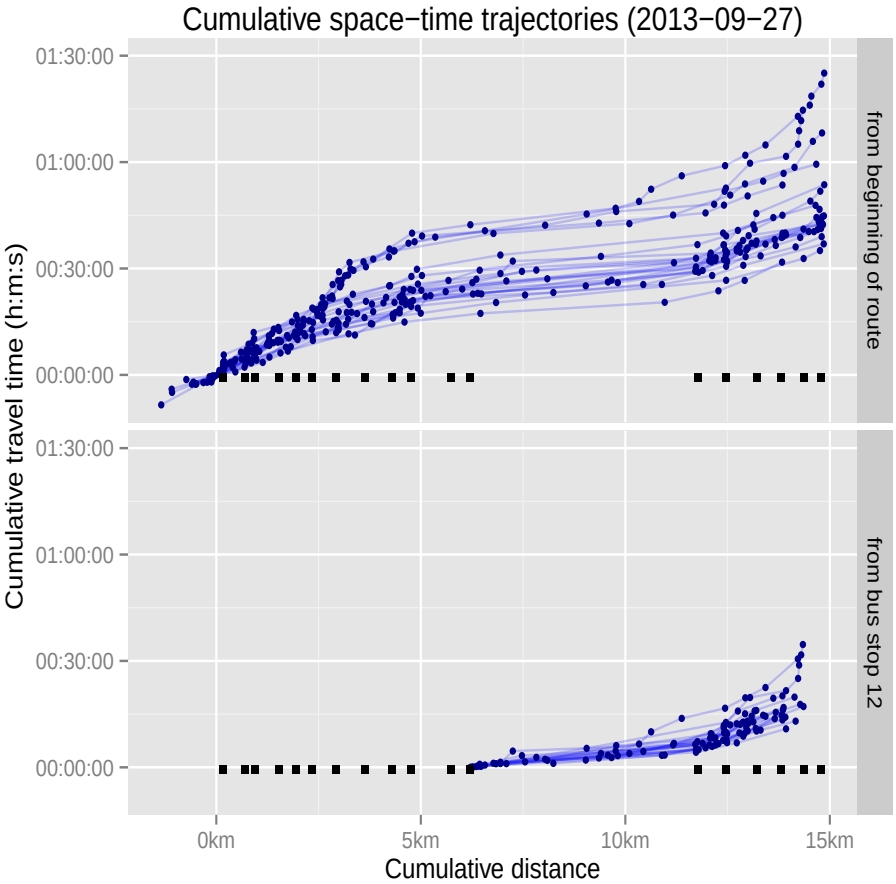
- **GPS (lat,lon,timestamp)**
  collected from public buses in the city of Rio de Janeiro.
- **Bus Route Data**
  provide piecewise linear representations of routes and bus stop locations.
- **Normalization**
  Map GPS coordinates onto a 1-dimensional scale measuring distance from origin.
- **Travel Times**
  inferred by differences between consecutive time stamps.

Cumulative Space-Time Trajectories



- GPS mapped onto a cumulative distance scale (x-axis).
- Cumulative travel time (y-axis) defined as zero at origin.
- Interpolation is required to infer time "zero".
- Scatter plots show raw measurements observed at irregular spatial locations.

Mathematical Notation/Setup

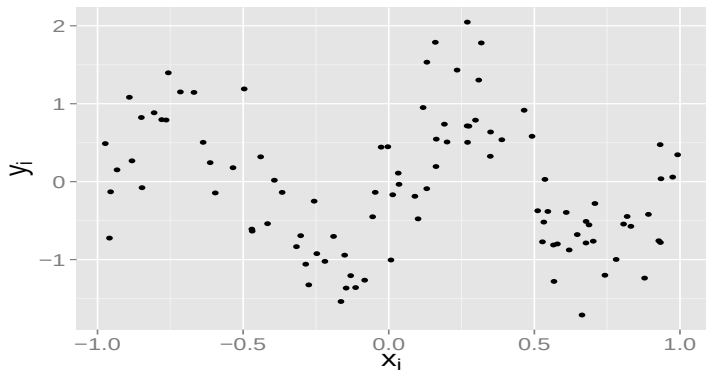


- Let $0 = p_0 < p_1 < \cdots < p_K$ denote the bus stops.
- Cumulative space-time trajectories normalized at 0 consist of cumulative distances $0 \leq dist_{ij} \leq p_K$, and cumulative travel times T_{ij} (bus i, obs j).
- To make future predictions for bus located at p_k we normalize trajectories at p_k (with $dist = 0$ and $T = 0$ at p_k).

Background - Spline Smoothing

Consider the simple scatterplot smoothing model

$$y_i = f(x_i) + \varepsilon_i.$$



Represent the function as

$$f(x) = \sum_{j=1}^{q} \beta_j \phi_j(x),$$

where $\phi_j(x)$ are known basis functions and β_j are coefficients to be estimated. An intuitive example is the piecewise linear representation, involving basis functions

$$\phi_1(x) = 1, \quad \phi_2(x) = x, \quad \phi_{j+2}(x) = (x - \tau_j)_+,$$

where $(x - \tau_j)_+ = \max(0, x - \tau_j)$, and τ_j are called knots.

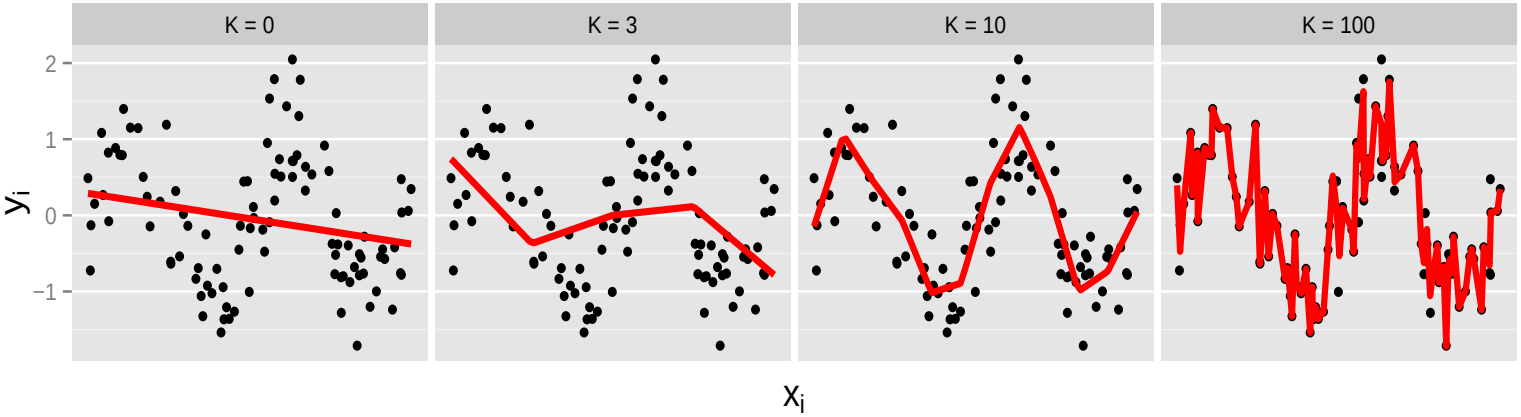
Piecewise linear spline functions

By setting up the design matrix appropriately the spline smoothing becomes a linear model:

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_1 & (x_1-\tau_1)_+ & \cdots & (x_1-\tau_K)_+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n-\tau_1)_+ & \cdots & (x_n-\tau_K)_+ \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}}_{\beta}.$$

Estimated by minimizing the **least squares**:

$$\min_{\beta} \|y - \mathbf{X}\beta\|^2.$$



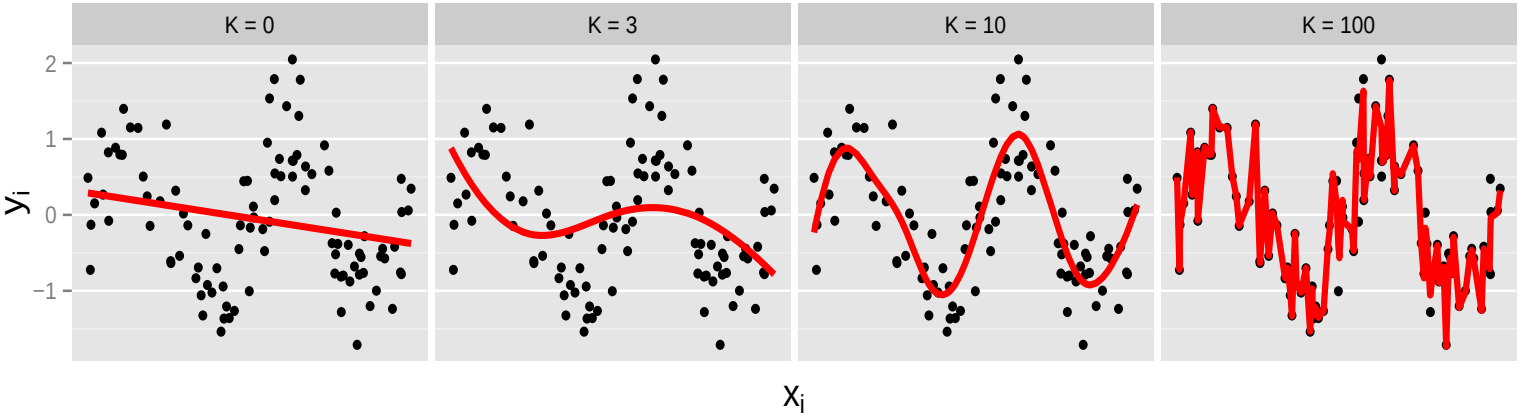
Cubic regression spline functions

By replacing the piecewise linear term with a cubic term in the design matrix, the model becomes:

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_1 & |x_1 - \tau_1|^3 & \cdots & |x_1 - \tau_K|^3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & |x_n - \tau_1|^3 & \cdots & |x_n - \tau_K|^3 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}}_{\beta}.$$

Estimated by minimizing the **least squares**:

$$\min_{\beta} \|y - \mathbf{X}\beta\|^2.$$



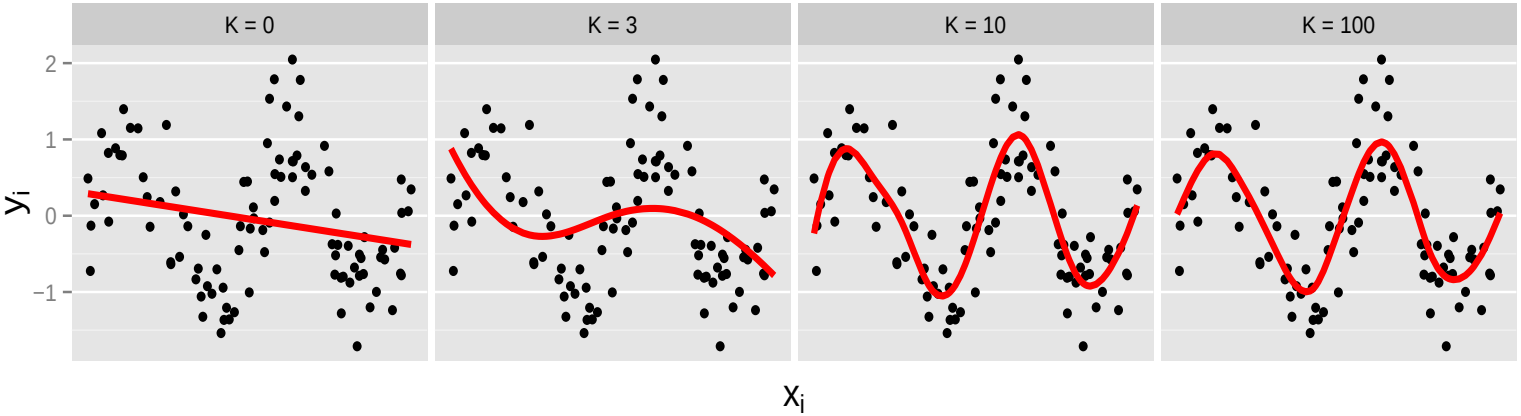
Penalized spline smoothing

By introducing a smoothness parameter λ we can control the smoothness of the model

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_1 & |x_1-\tau_1|^3 & \cdots & |x_1-\tau_K|^3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & |x_n-\tau_1|^3 & \cdots & |x_n-\tau_K|^3 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}}_{\beta}.$$

We minimize the **penalized least squares** and estimate λ by Generalized Cross Validation (GCV):

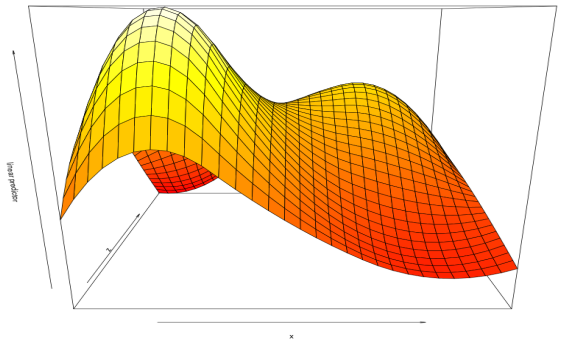
$$\min_{\beta} \|y - \mathbf{X}\beta\|^2 + \lambda\beta' D\beta, \quad D = diag(0, 0, 1, \ldots, 1)$$



Penalized tensor product smoothing

Consider the bivariate smoothing model

$$y_i = f(x_{1i}, x_{2i}) + \varepsilon_i.$$



Represent the function as

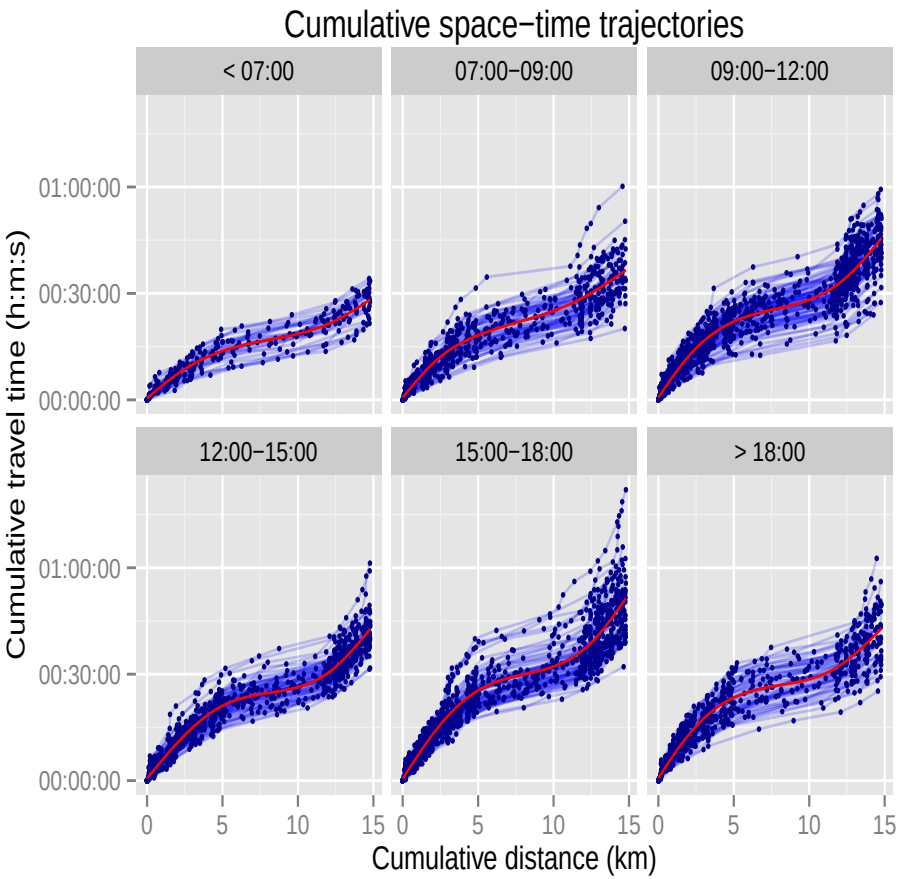
$$f(x_1, x_2) = \sum_{j=1}^{q_1} \sum_{k=1}^{q_2} \beta_{jk} \phi_j(x_1) \psi_k(x_2),$$

where $\phi_j(x)$ and $\psi_k(x)$ are known basis functions (e.g. linear/cubic splines) and β_{jk} are coefficients to be estimated.

By appropriately specifying the design matrix in terms of the basis functions the above model becomes linear:

$$y = \mathbf{X}\beta + \varepsilon.$$

Proposed solution - Additive Models



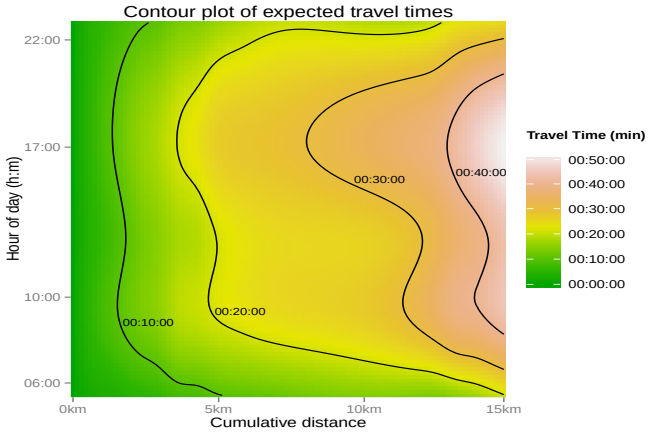
- ... allow the linear predictor to depend on unknown smooth functions of predictor variables.
- (Left): Trajectories of route 121 stratified by hour.
- Note how travel time changes smoothly as function of distance.
- Note how functional relationship changes with the hour, $time_i$.

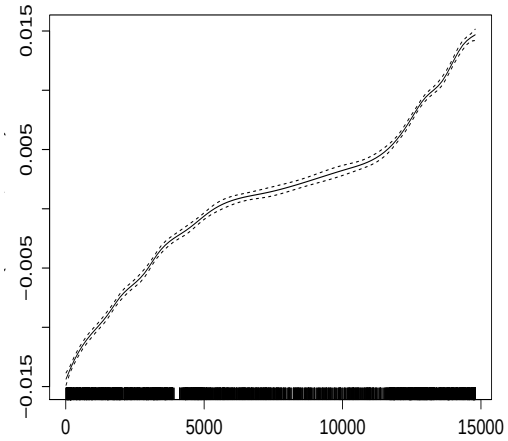
Model 1: Basic Additive Model (BAM)

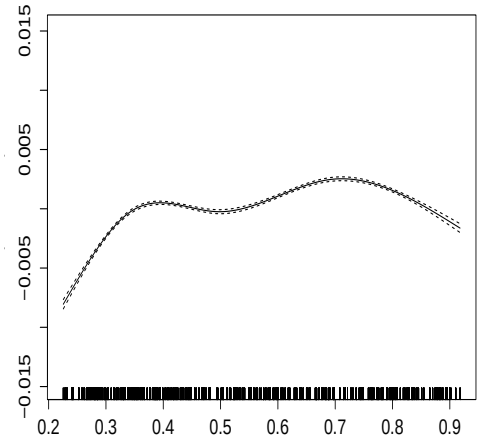
$$T_{ij} = \beta_0 + f_1(dist_{ij}) + f_2(time_i) + f_3(dist_{ij}, time_i) + \varepsilon_{ij}.$$

Estimation: Represent f_1, f_2 with penalized splines and f_3 with penalized tensor product. Model becomes linear in parameters.

Analysis of route 121 (Copacabana-Centro)







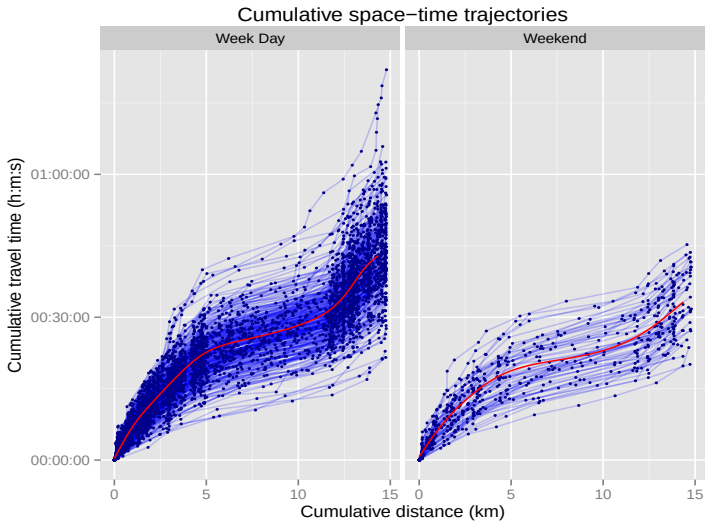
Model 1: Basic Additive Model (BAM)

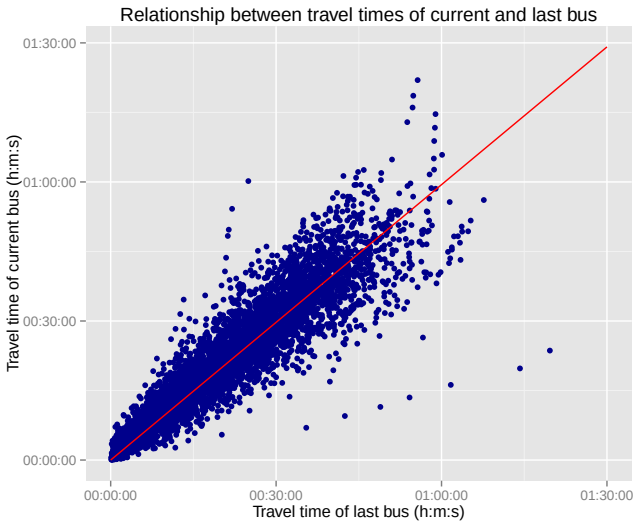
$$T_{ij} = \beta_0 + f_1(dist_{ij}) + f_2(time_i) + f_3(dist_{ij}, time_i) + \varepsilon_{ij},$$

Summary: All terms are significant and adjusted $\underline{R^2 = 0.903}$.

Extended Additive Model

Additive Models are flexible and allow for additional linear predictors.





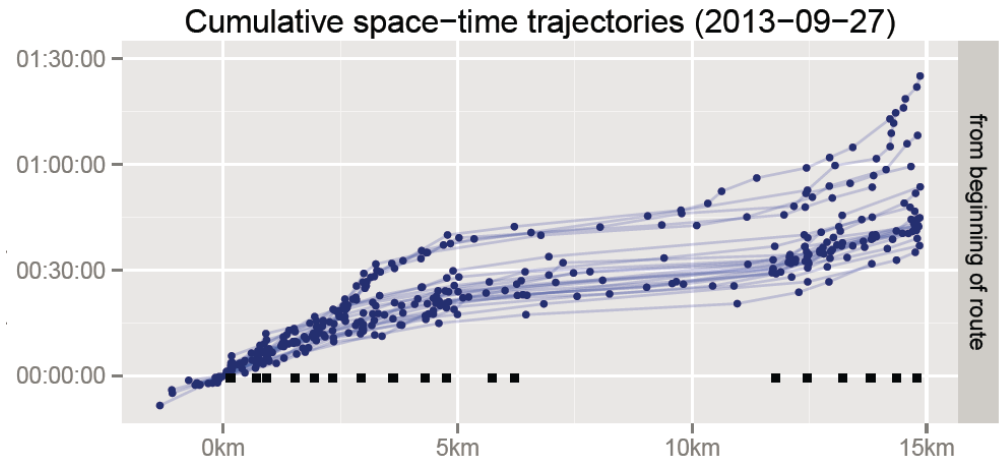
Model 2: Extended Additive Model (EAM)

$$T_{ij} = \beta_0 + \beta_1 \cdot weekend_i + f_1(dist_{ij}, weekend_i) + \beta_2 \cdot T_{ij}^{\text{last}} + f_2(time_i) + f_3(dist_{ij}, time_i) + \varepsilon_{ij}.$$

Summary: All terms are significant and adjusted $\underline{R^2 = 0.919}$.

Additive Mixed Model

Recall that time zero for bus i was inferred by interpolating two consecutive time stamps before and after origin.



Model 3: Additive Mixed Model (AMM)

$$T_{ij} = \beta_0 + b_{0i} + \beta_1 \cdot weekend_i + f_1(dist_{ij}, weekend_i)$$
$$+ \beta_2 \cdot T_{ij}^{\text{last}} + f_2(time_i) + f_3(dist_{ij}, time_i) + \varepsilon_{ij},$$

where $b_{0i} \sim N(0, \sigma_b^2)$ is a corrective random intercept.

Summary: $\hat{\sigma}_b = 3$ minutes and adjusted $\underline{R^2 = 0.968}$.

Experiments

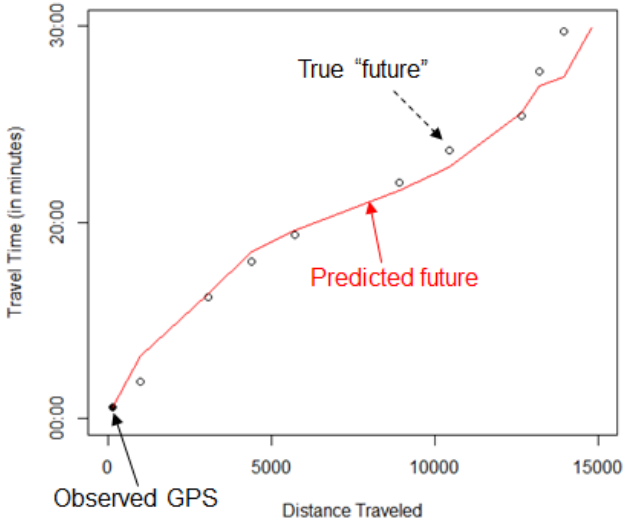
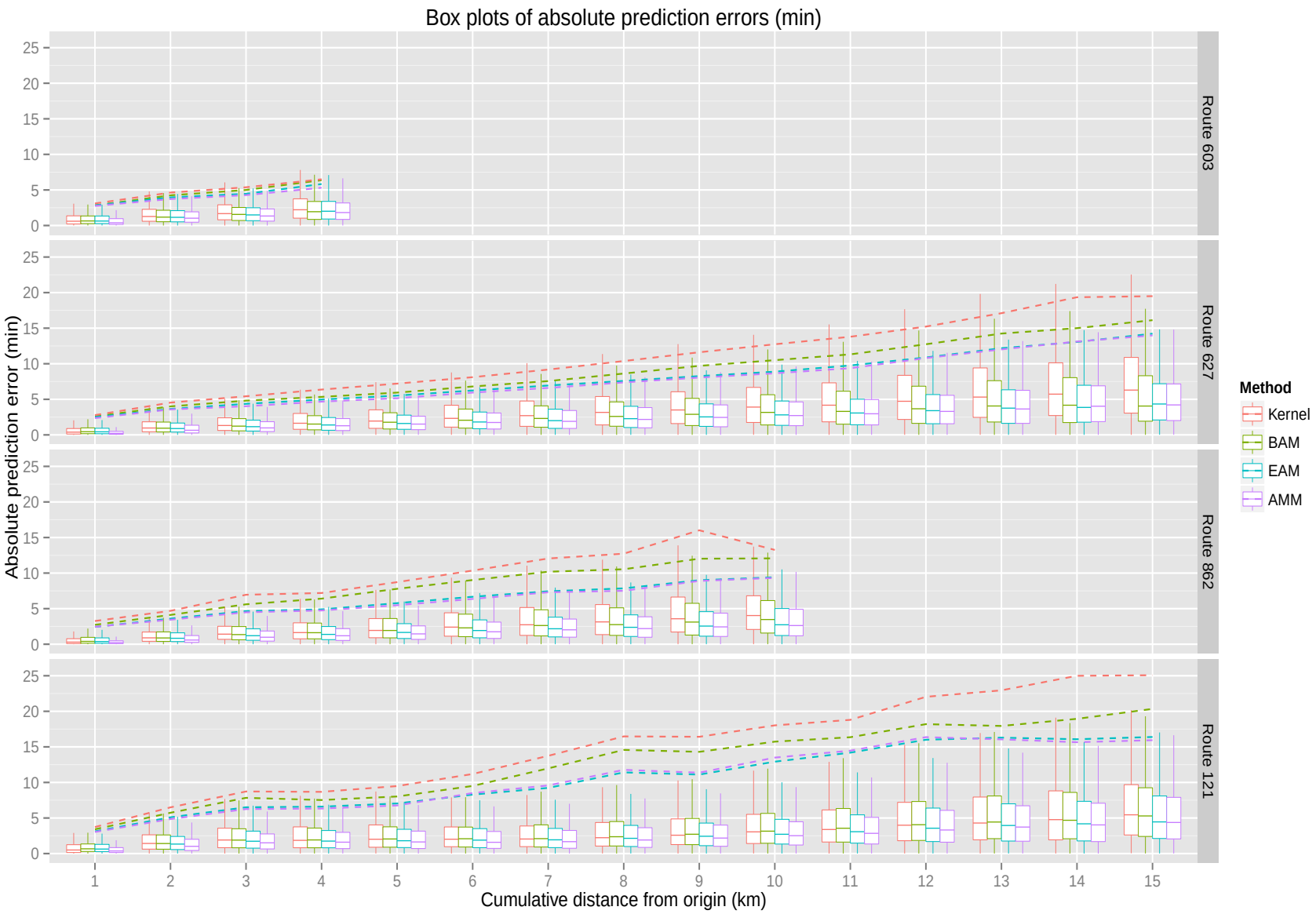


Table : Route data summary

Route	# trajectories	# stops	Length
603	1,276	15	4km
627	1,325	54	15km
862	7,882	24	10km
121	2,515	18	15km

- **Data:** Four bus routes: 603, 627, 862, and 121.
- **Test data:** 14 random days (out of 3 months of data).
- **Training data:** 10, 20, 30 days before each test date.
- **Prediction:** made for all buses i in test set from every bus stop p_k until end of route.
- **Error:** calculated for bus i at observed $dist_{ij}$ and stratified by prediction distance $|dist_{ij} - p_k|$.

Experimental results



Experimental results

Table : Mean Absolute Relative Error (MARE)

		Method				
Route	# days	BAM	EAM	AMM	Kernel	SVM
603	10	19.9%	19.7%	**18.4%**	21.3%	64.4%
	20	20.1%	19.8%	**18.5%**	21.3%	64.7%
	30	19.8%	19.6%	**18.3%**	21.3%	64.8%
627	10	16.3%	14.7%	**13.8%**	18.1%	28.8%
	20	15.2%	14.2%	**13.4%**	17.3%	30.0%
	30	15.1%	14.0%	**13.2%**	17.1%	29.4%
862	10	22.1%	19.5%	**18.0%**	23.8%	26.4%
	20	22.5%	19.3%	**18.0%**	23.6%	26.8%
	30	22.2%	19.3%	**17.9%**	23.4%	25.6%
121	10	23.1%	20.9%	**19.2%**	23.9%	41.5%
	20	22.9%	20.7%	**19.1%**	23.6%	41.4%
	30	22.7%	20.3%	**18.9%**	23.4%	41.2%

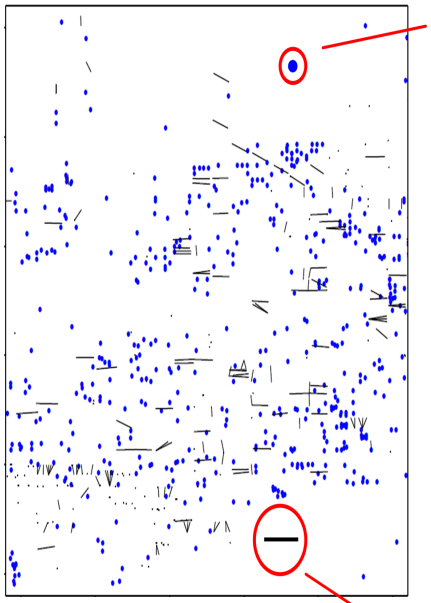
$$\text{MARE} = (1/N) \sum_{ij} |T_{ij} - \hat{T}_{ij}|/T_{ij}$$

Conclusions

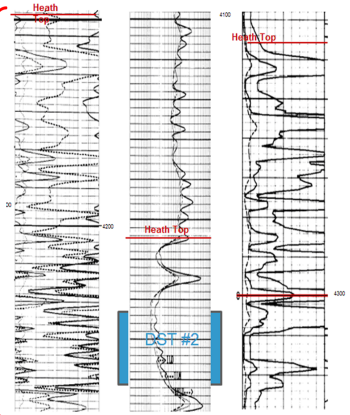
- Proposed solution:
  - models travel times directly using raw irregular GPS data.
  - models spatial and temporal effects through smooth functions thus avoiding any discretization.
  - allows for flexible incorporation of additional predictors.
- We showed that by including a random intercept we correct for an interpolation error.
- Demonstrated on a large real-world GPS data that our method achieved superior performance (as compared to existing methods).

Other Projects at IBM - Oil&Gas

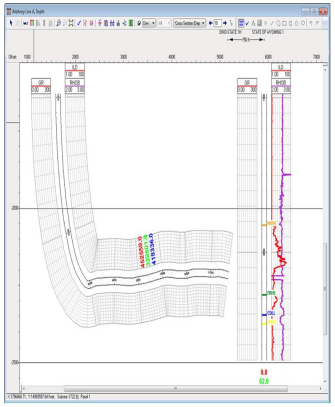
Field with well logs from vertical wells and production data from horizontal wells



Well log data



Horizontal well production data



Model Building

Functional Principal Component Analysis

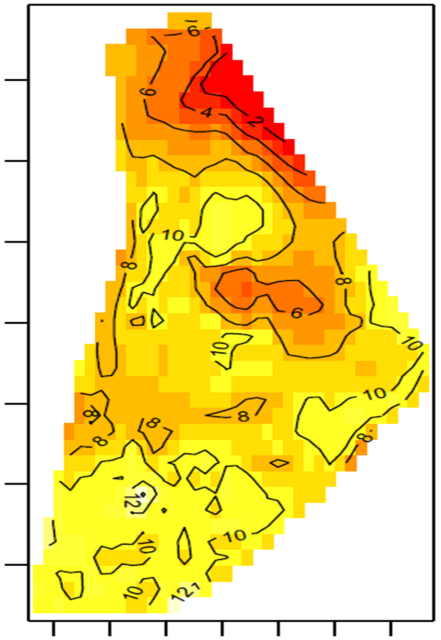
↓

Kriging

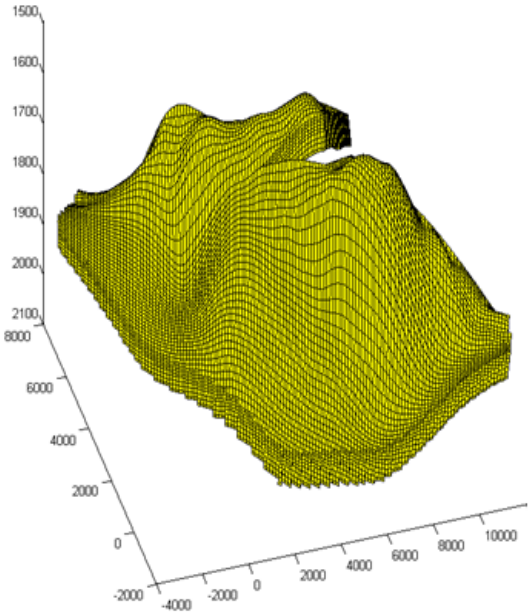
↓

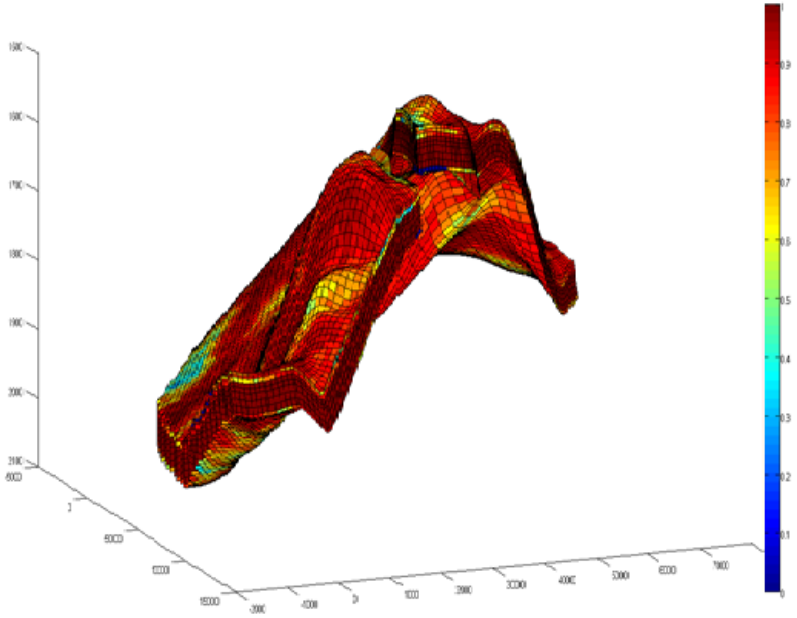
Multiple Linear Regression

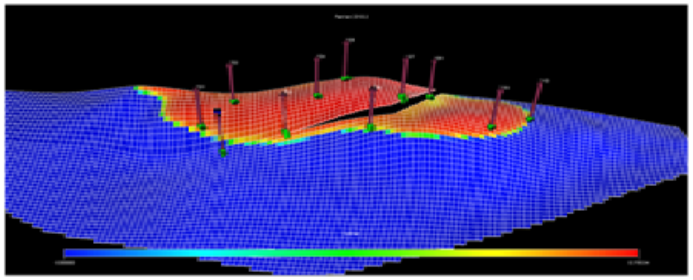
Prediction (color maps)



Other Projects at IBM - Oil&Gas







Publications/Contact

Kormaksson, M., Barbosa, L., Vieira, M., and Zadrozny, B. (2014) *"Bus Travel Time Predictions Using Additive Models."* Proc. of 14th IEEE Int'l Conference on Data Mining (ICDM).

Barbosa, L., Kormaksson, M., Vieira, M., Lage, R., and Zadrozny, B. (2014) *"Vistradas: Visual Analytics for Urban Trajectory Data."* In Proceedings of the 15th Brazilian Symposium on Geoinformatics (GeoInfo).

Kormaksson, M., Vieira, M., and Zadrozny, B. (2015) *"A Data Driven Method For Sweet Spot Identification In Shale Plays Using Well Log Data."* SPE Digital Energy Conference.

Contact: matkorm@br.ibm.com