Crowdsourced mobility data: techniques and applications

Francisco C. Pereira

With: Filipe Rodrigues, Kristian Henrickson, Moshe Ben-Akiva, Yang Lu, Haizheng Zhang, Constantinos Antoniou

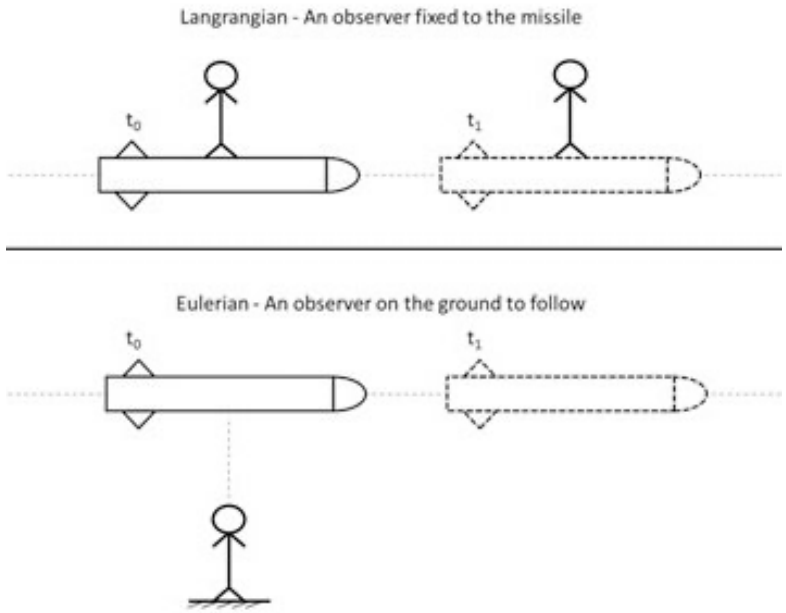
Technical University of Denmark DTU

MLSM
Machine Learning for Smart Mobility group
http://mlsm.man.dtu.dk

Outline

- Two (opposing?) modelling paradigms
- Machine Learning
  - Imputation with multi-output modeling
  - Uncertainty analysis with heteorskedastic modeling
- Simulation
  - Online calibration

Two (opposing?) paradigms



Two (opposing?) paradigms

Simulation
- Model phenomenon
- First principles/domain rules
- Causal relations

Machine Learning
- Approximation of input/output
- Statistical theory, information theory
- Correlations/patterns

Machine Learning

- Imputation with multi-output modeling
- Robust prediction with uncertainty modeling

DTU MLSM
Machine Learning for Smart Mobility

Imputation with multi-output modeling

- Problem statement
  - For its nature, crowdsourced data is prone to imperfect observations
  - Missing data, noise, unbalanced observations

Imputation with multi-output modeling

- Context
  - Google research dataset
  - Android and iPhone Gmaps users
  - Speeds and flows (buckets) per 13 network links in 4 areas
  - Aggregations by 5 min intervals

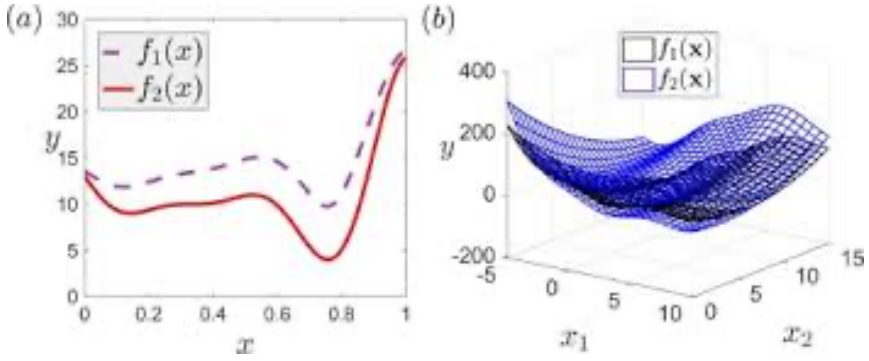
TABLE I
DESCRIPTIVE STATISTICS OF STUDY AREAS.

Area	Num. Lanes	Road type	Traffic lights	Speed limit	Average speed	Speed std. dev.
1	2 (+1 Bus)	Arterial with BRT	100m before	50 km/h	23.3 km/h	16.9 km/h
2	2 (+1 Bus)	Arterial with BRT	100m after	50 km/h	30.3 km/h	18.3 km/h
3	2	Local road	150m after	50 km/h	21.8 km/h	16.3 km/h
4	2 (+1 Bus)	Intersection	Just before	50 km/h	29.0 km/h	15.5 km/h

DTU MLSM
Machine Learning for Smart Mobility

Approach

- Jointly model the entire “network” in a multi-output model
- Principle is that correlation patterns (over a large enough sample) make up for real joint distribution
- We model this as a multi-output Gaussian Processes model



DTU MLSM
Machine Learning for Smart Mobility

Experimental design

- 6 months of data
- Assume existing non-missing data is “perfect” (to use it as ground truth)
- Random selection of missing data (50% and 75%) to “hide” from the model
- Check how model approximates the ground-truth

Results

TABLE II
RESULTS FOR DIFFERENT METHODS ACROSS STUDY AREA 1 AND 2 FOR A MISSING RATIO OF 50%.

Area	Method	MAE	RMSE	RAE	R^2
Area 1	Naive	4.230	6.843	72.856	0.308
	ARIMA	2.874	5.160	49.490	0.606
	Lin. Reg. (B)	2.492	4.369	42.911	0.718
	Lin. Reg. (A)	2.486	4.361	42.823	0.719
	Lin. Reg. (B+A)	2.442	4.294	42.053	0.727
	kNN (B)	2.836	4.631	48.835	0.683
	kNN (A)	2.817	4.599	48.510	0.687
	kNN (B+A)	2.722	4.446	46.874	0.708
	B-EM (B)	3.077	5.116	52.994	0.613
	B-EM (A)	3.058	5.069	52.671	0.620
	B-EM (B+A)	2.713	4.641	46.730	0.682
	PPCA (B)	3.132	5.133	53.939	0.611
	PPCA (A)	3.107	5.143	53.508	0.609
	PPCA (B+A)	2.838	4.728	48.871	0.670
	Bi-LSTM (B)	2.616	4.502	45.054	0.700
	Bi-LSTM (A)	2.595	4.513	44.693	0.699
	Bi-LSTM (B+A)	2.583	4.495	44.479	0.701
	Indep. GP	2.365	4.202	40.740	0.739
	VARMA (B)	2.395	4.248	41.242	0.733
	VARMA (A)	2.395	4.248	41.244	0.733
	VARMA (B+A)	2.377	4.217	40.946	0.737
	Multi-GP (B)	1.900	3.599	32.719	0.809
	Multi-GP (A)	**1.864**	**3.560**	**32.102**	**0.813**
	Multi-GP (B+A)	1.951	3.786	33.604	0.788
Area 2	Naive	4.460	7.513	66.120	0.389
	ARIMA	3.770	6.647	55.890	0.522
	Lin. Reg. (B)	2.631	4.846	39.013	0.746
	Lin. Reg. (A)	2.619	4.828	38.826	0.748
	Lin. Reg. (B+A)	2.582	4.761	38.275	0.755
	kNN (B)	3.060	5.216	45.361	0.706
	kNN (A)	3.022	5.117	44.800	0.717
	kNN (B+A)	2.981	5.067	44.200	0.722
	B-EM (B)	3.552	5.990	52.654	0.612
	B-EM (A)	3.509	5.872	52.019	0.627
	B-EM (B+A)	3.228	5.546	47.861	0.667
	PPCA (B)	3.752	6.314	55.619	0.569
	PPCA (A)	3.636	6.074	53.903	0.601
	PPCA (B+A)	3.421	5.640	50.712	0.656
	Bi-LSTM (B)	2.729	4.934	40.459	0.737
	Bi-LSTM (A)	2.706	4.895	40.122	0.741
	Bi-LSTM (B+A)	2.704	4.852	40.090	0.745
	Indep. GP	2.482	4.611	36.795	0.770
	VARMA (B)	2.538	4.693	37.688	0.761
	VARMA (A)	2.530	4.641	37.510	0.767
	VARMA (B+A)	2.511	4.609	37.221	0.770
	Multi-GP (B)	2.167	4.177	32.132	0.811
	Multi-GP (A)	**2.104**	**4.063**	**31.199**	**0.821**
	Multi-GP (B+A)	2.105	4.071	32.201	0.820

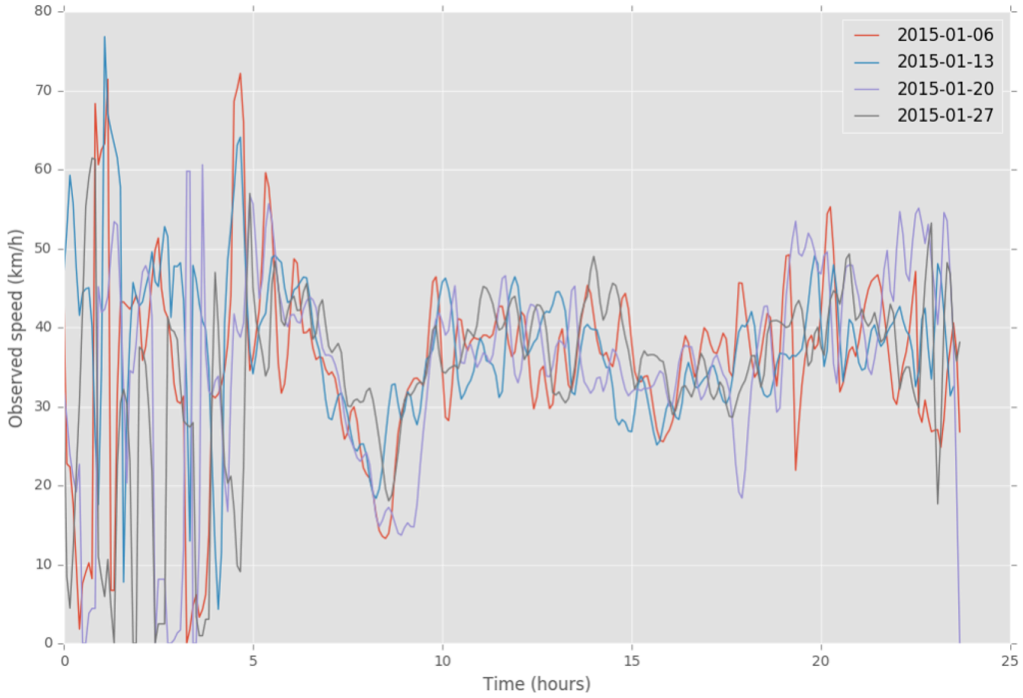
TABLE III
RESULTS FOR DIFFERENT METHODS ACROSS STUDY AREA 3 AND 4 FOR MISSING RATIO OF 50%.

Area	Method	MAE	RMSE	RAE	R^2
Area 3	Naive	3.839	6.139	69.956	0.313
	ARIMA	2.147	3.741	39.124	0.745
	Lin. Reg. (B)	2.223	3.890	40.504	0.724
	Lin. Reg. (A)	2.204	3.864	40.153	0.728
	Lin. Reg. (B+A)	2.170	3.806	39.548	0.736
	kNN (B)	2.489	4.039	45.360	0.703
	kNN (A)	2.483	4.018	45.239	0.706
	kNN (B+A)	2.400	3.906	43.735	0.722
	B-EM (B)	2.601	4.289	47.403	0.665
	B-EM (A)	2.568	4.207	46.800	0.677
	B-EM (B+A)	2.259	3.802	41.160	0.736
	PPCA (B)	2.689	4.349	48.994	0.655
	PPCA (A)	2.675	4.301	48.748	0.663
	PPCA (B+A)	2.428	3.971	44.244	0.712
	Bi-LSTM (B)	2.348	4.046	42.778	0.702
	Bi-LSTM (A)	2.325	3.998	42.372	0.709
	Bi-LSTM (B+A)	2.289	3.965	41.702	0.713
	Indep. GP	2.075	3.717	37.817	0.748
	VARMA (B)	2.085	3.698	37.998	0.751
	VARMA (A)	2.076	3.686	37.827	0.752
	VARMA (B+A)	2.061	3.660	37.553	0.756
	Multi-GP (B)	1.594	2.972	29.037	0.839
	Multi-GP (A)	1.585	**2.962**	28.886	**0.840**
	Multi-GP (B+A)	**1.579**	2.969	**28.781**	0.839
Area 4	Naive	3.226	5.314	58.934	0.499
	ARIMA	1.971	3.466	36.008	0.787
	Lin. Reg. (BR)	1.909	3.533	34.876	0.778
	Lin. Reg. (BL)	1.835	3.426	33.528	0.792
	Lin. Reg. (A)	1.810	3.370	33.067	0.798
	Lin. Reg. (BL+A)	1.759	3.296	32.136	0.807
	kNN (BR)	2.531	4.146	46.233	0.695
	kNN (BL)	2.124	3.633	38.805	0.766
	kNN (A)	2.093	3.587	38.234	0.772
	kNN (BL+A)	2.031	3.484	37.101	0.784
	EM (BR)	2.254	3.878	39.124	0.721
	EM (BL)	2.123	3.775	38.776	0.747
	EM (A)	2.037	3.590	37.206	0.771
	EM (BL+A)	1.740	3.221	31.790	0.816
	PPCA (BR)	2.212	3.876	40.563	0.729
	PPCA (BL)	2.217	3.882	40.497	0.732
	PPCA (A)	2.103	3.617	38.417	0.768
	PPCA (BL+A)	1.933	3.375	35.318	0.798
	Bi-LSTM (BR)	1.966	3.578	35.988	0.771
	Bi-LSTM (BL)	1.951	3.534	35.641	0.778
	Bi-LSTM (A)	1.914	3.472	34.961	0.786
	Bi-LSTM (BL+A)	1.854	3.376	33.864	0.798
	Indep. GP	1.749	3.275	31.947	0.810
	VARMA (BL)	1.780	3.310	32.525	0.805
	VARMA (BR)	1.790	3.329	32.708	0.803
	VARMA (A)	1.776	3.305	32.450	0.806
	VARMA (BL+A)	1.770	3.296	32.332	0.807
	Multi-GP (BR)	1.660	3.166	30.331	0.822
	Multi-GP (BL)	1.256	2.630	22.937	0.877
	Multi-GP (A)	1.191	**2.477**	21.758	**0.891**
	Multi-GP (BL+A)	**1.190**	2.486	**21.746**	0.890

DTU MLSM
Machine Learning for Smart Mobility

Uncertainty analysis with heteroskedastic modeling

- Problem statement:
  - Assumption of constant noise variance is often too simplistic
    - Can bias mean estimation
    - Gives wrong notion of uncertainty (e.g. confidence intervals)



DTU MLSM
Machine Learning for Smart Mobility

Uncertainty analysis with heteroskedastic modeling

- Google research dataset (as before)
- Model in two components

$$y_t = f(t) + \epsilon_t,$$

$$p(\mathbf{f}) = \mathcal{GP}(m_f(t) = 0, k_f(t, t'))$$

$$\epsilon_t \sim \mathcal{N}(\epsilon_t | 0, r(t)) \rightarrow r(t) = e^{g(t)} \rightarrow g(t) \sim \mathcal{GP}(\mu_0, k_g(t, t'))$$

- Condition noise on time (and flow)

Results

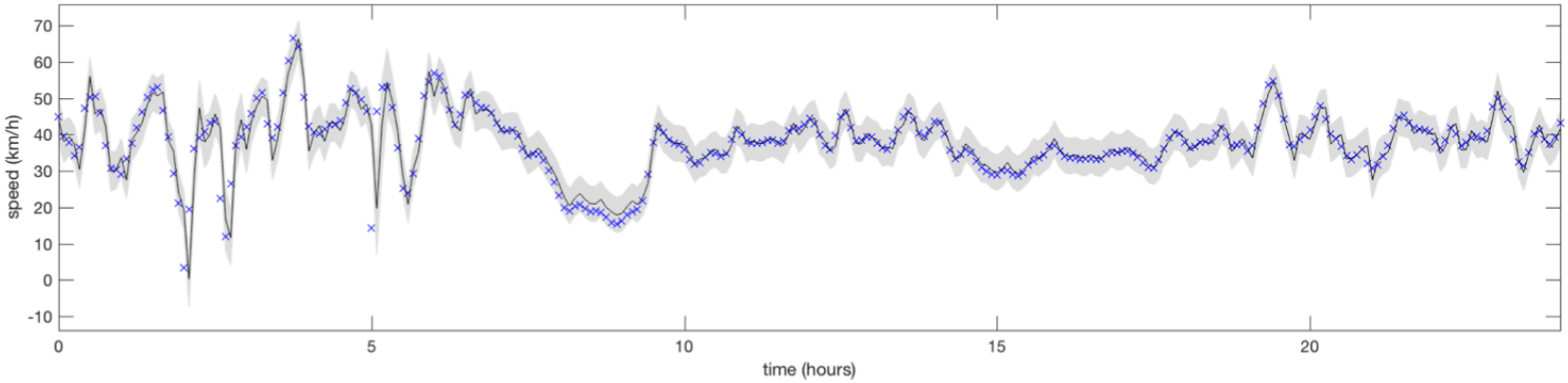
- Does it improve the mean itself?

Place ID	Method	MAE	RAE	R^2
1	Lin. Interp.	2.840	40.897	0.717
	GP	2.550	37.244	0.752
	HGP	2.460	35.917	0.757
	SSRC-HGP	**2.414**	**34.764**	**0.759**
2	Lin. Interp.	2.869	40.526	0.732
	GP	2.570	36.186	0.760
	HGP	**2.440**	**34.361**	**0.775**
	SSRC-HGP	2.466	34.837	0.771
3	Lin. Interp.	2.839	51.547	0.617
	GP	2.477	44.983	0.672
	HGP	2.378	43.177	0.677
	SSRC-HGP	**2.336**	**42.412**	**0.694**
4	Lin. Interp.	2.902	43.292	0.717
	GP	2.592	38.678	0.755
	HGP	2.487	37.112	0.756
	SSRC-HGP	**2.464**	**36.766**	**0.757**
5	Lin. Interp.	2.366	43.446	0.714
	GP	2.073	38.073	0.759
	HGP	**1.986**	**36.481**	**0.765**
	SSRC-HGP	2.020	37.112	0.755
6	Lin. Interp.	4.593	50.897	0.620
	GP	4.103	45.471	0.671
	HGP	4.122	45.685	0.657
	SSRC-HGP	**3.945**	**43.724**	**0.690**
7	Lin. Interp.	2.761	47.657	0.650
	GP	2.479	42.799	0.693
	HGP	2.322	40.094	0.709
	SSRC-HGP	**2.232**	**39.258**	**0.714**

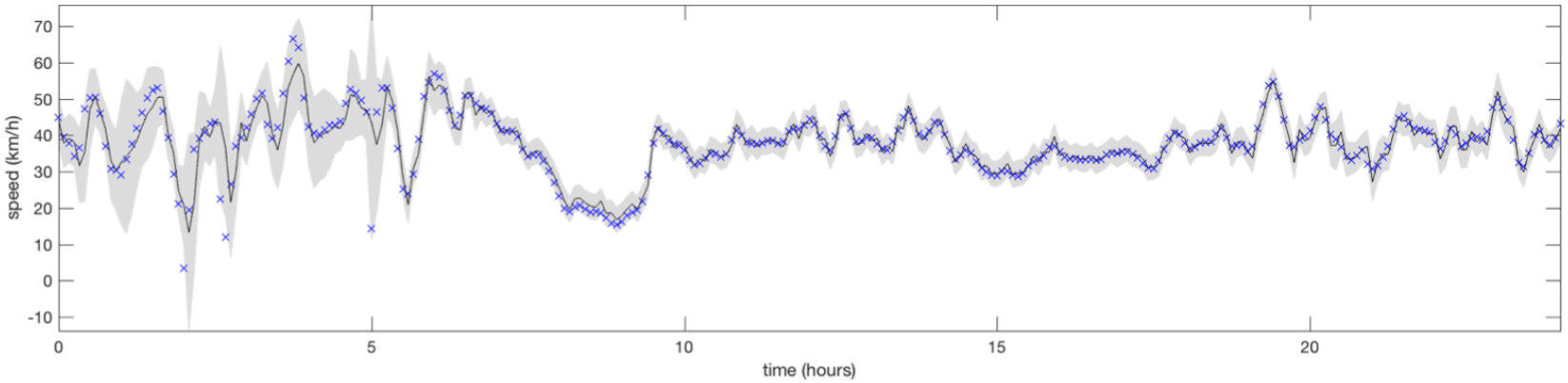
- What about the intervals?

		Eval: all periods				Eval: day periods		
Place ID	Method	NLPD	ICP	MIL	RMIL	ICP	MIL	RMIL
1	ARIMA	0.796	0.949	18.257	53.787	0.994	18.264	65.089
	GP	0.679	0.961	18.193	52.555	0.996	18.213	60.978
	HGP	0.453	0.950	14.987	48.268	0.995	14.506	55.857
	SSRC-HGP	**0.049**	0.955	11.062	32.179	0.987	8.644	35.123
2	ARIMA	0.674	0.959	17.761	50.249	0.995	17.760	53.118
	GP	0.653	0.962	17.909	43.940	0.995	17.897	51.152
	HGP	0.416	0.948	13.564	45.110	0.990	13.285	53.585
	SSRC-HGP	**0.082**	0.952	11.086	33.840	0.981	9.019	35.163
3	ARIMA	0.766	0.949	17.458	49.346	0.995	17.464	61.664
	GP	0.648	0.959	17.311	49.636	0.996	17.369	59.886
	HGP	0.421	0.953	14.535	49.692	0.995	14.208	59.244
	SSRC-HGP	**0.025**	0.955	10.755	34.508	0.987	8.381	39.421
4	ARIMA	0.783	0.942	18.017	50.174	0.994	18.023	59.777
	GP	0.728	0.960	18.114	51.081	0.993	20.263	48.937
	HGP	0.658	0.929	11.899	44.970	0.989	13.541	46.938
	SSRC-HGP	**0.116**	0.955	12.375	35.668	0.983	9.497	35.189
5	ARIMA	0.766	0.946	17.458	49.346	0.995	17.464	61.664
	GP	0.646	0.959	17.293	51.373	0.996	17.316	58.798
	HGP	0.418	0.953	14.514	49.029	0.995	14.255	60.422
	SSRC-HGP	**0.030**	0.955	10.778	35.714	0.987	8.409	39.845
6	ARIMA	0.766	0.945	17.458	49.346	0.995	17.464	61.664
	GP	0.649	0.959	17.344	51.734	0.996	17.596	64.397
	HGP	0.420	0.954	14.669	48.534	0.994	14.448	60.189
	SSRC-HGP	**0.043**	0.954	10.865	35.615	0.986	8.400	35.714
7	ARIMA	0.766	0.946	17.458	49.345	0.994	17.467	61.668
	GP	0.646	0.959	17.281	49.925	0.996	17.307	61.216
	HGP	0.428	0.953	14.715	50.679	0.994	14.180	60.831
	SSRC-HGP	**0.040**	0.951	10.890	33.974	0.987	8.508	38.165

Results



(c) HGP



(d) SSRC-HGP

MLSM
Machine Learning for Smart Mobility

Simulation paradigm

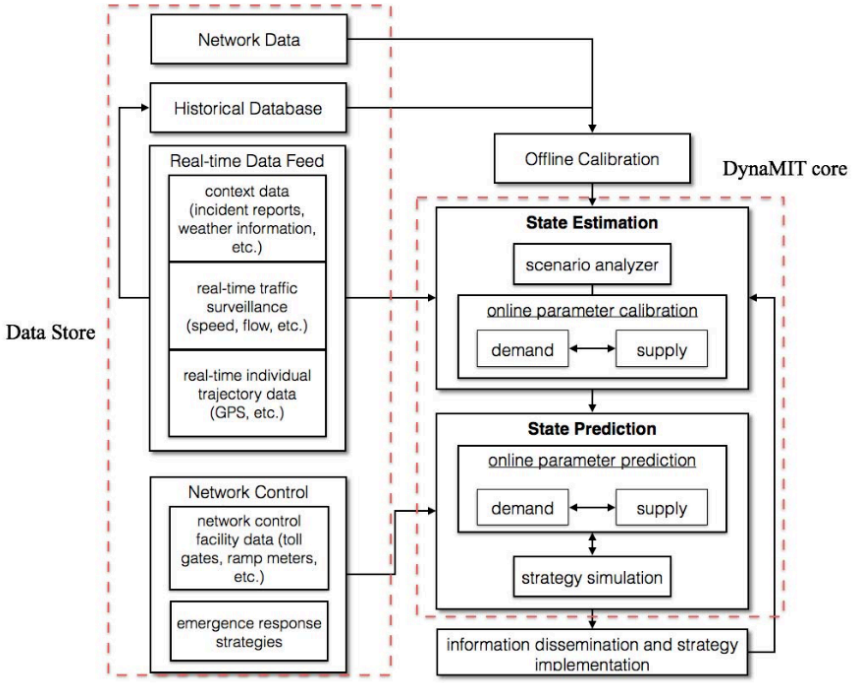


FIGURE 2 Architecture of DynaMIT2.0

- Extended Kalman Filter

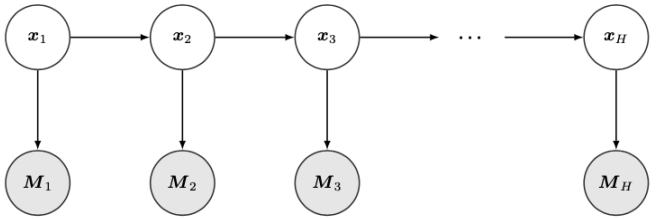


Fig. 1: State space model with measurements

Table 5: Flow RMSN for state estimation and predictions for 15:00-19:00

Experiment	Estimation	Prediction RMSN		
	RMSN	Step 1	Step 2	Step 3
CEKF(1)	13.5%	21.0%	26.2%	34.7%
CEKF(2)	9.8%	18.8%	24.2%	31.9%
CEKF(5)	10.8%	15.4%	19.3%	26.6%

DTU MLSM
Machine Learning for Smart Mobility

Simulation paradigm

- The point is…
  - The ML methods shown (and the **vast** majority of them) are correlational
  - They are excellent in pattern recognition, distribution learning
  - They fail with structural changes (e.g. network changes, behaviour change)
  - Simulation tools can address those limitations
  - **Combine two paradigms with causal ML!**

Content

- *Multi-output Gaussian processes for crowdsourced traffic data imputation: https://arxiv.org/pdf/1812.08739.pdf*
- *Heteroscedastic Gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data: https://arxiv.org/abs/1812.08733*
- *DynaMIT2.0: Architecture Design and Preliminary Results on Real-time Data Fusion for Traffic Prediction and Crisis Management: https://ieeexplore.ieee.org/document/7313455*