Missing Data In Traffic Estimation: A Variational Autoencoder Imputation Method

Guillem Boquet, Jose Lopez Vicario, Antoni Morell & Javier Serrano

Wireless Information Networking Group
Telecommunications and Systems Engineering Department
Universitat Autònoma de Barcelona (UAB)

ICASSP 2019, Brighton, UK

OUTLINE

Context

- Future Intelligent Transportation Systems (ITS)
- Road Traffic Forecast relevance
- Deep Learning trend

Major challenges

- of future road traffic forecast [Laña et al., 2018]:
    - **Quality of the data**
    - Network-level predictions
    - Spatiotemporal forecasts
    - Model selection techniques
    - Etc.

Missing data problem

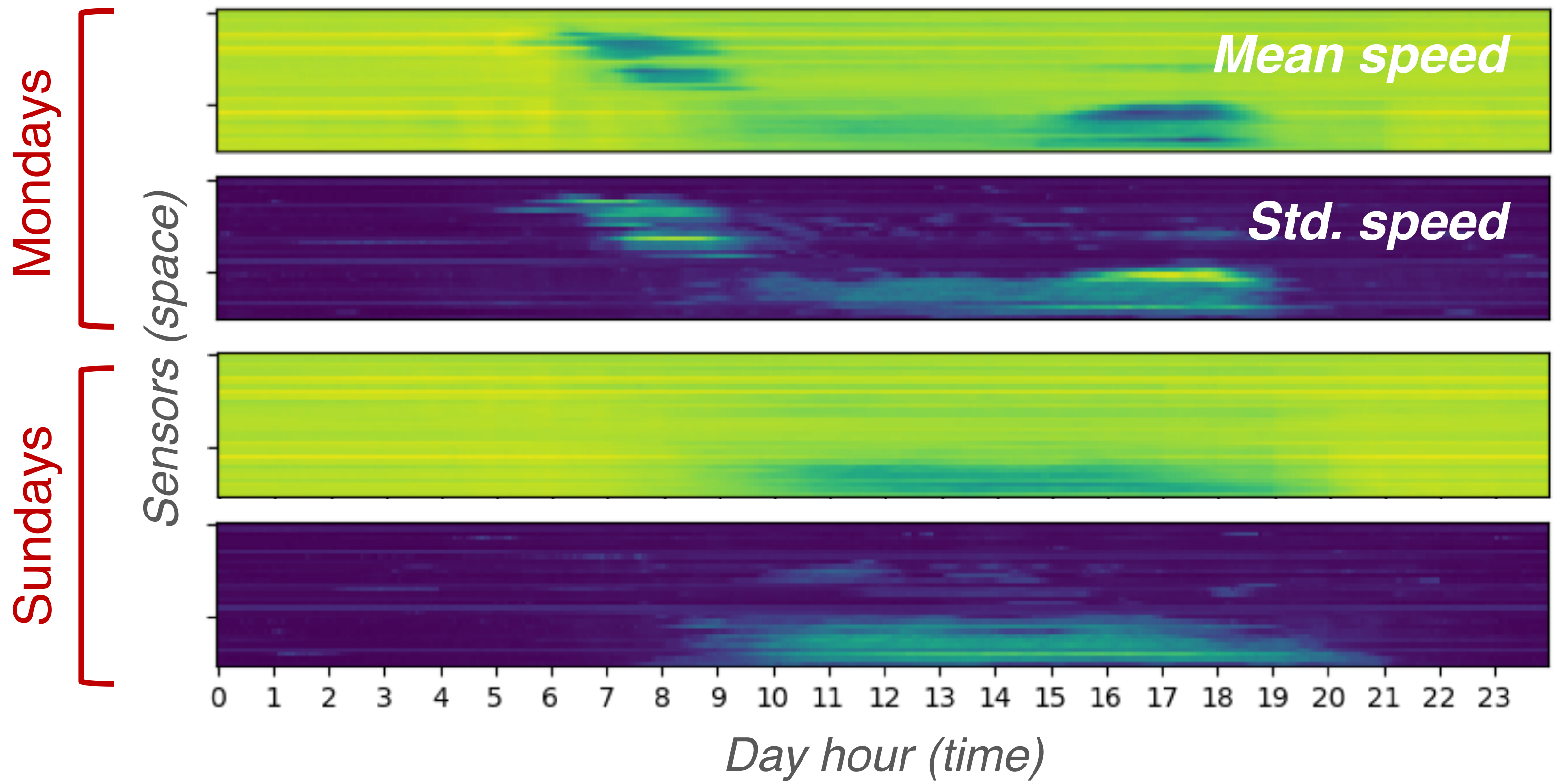
- All real-world traffic data sets contain missing values (MVs)
- Negatively affect estimation accuracy but often underestimated [Laña, 2018; Vlahogianni, 2014]
- Current imputation methods in traffic forecast:
  - ARIMA, KNN and PCA based methods
  - Automated clustering tool [Laña et al., 2018-b]
  - LSTM, SVR and collaborative filtering [Li et al., 2018]
  - Bayesian tensor decomposition model [Chen et al., 2019]

Proposal

Assumption:

- Traffic data samples are not randomly generated

Traffic patterns:



Proposal

- Exists a non-linear latent manifold from which traffic data are generated

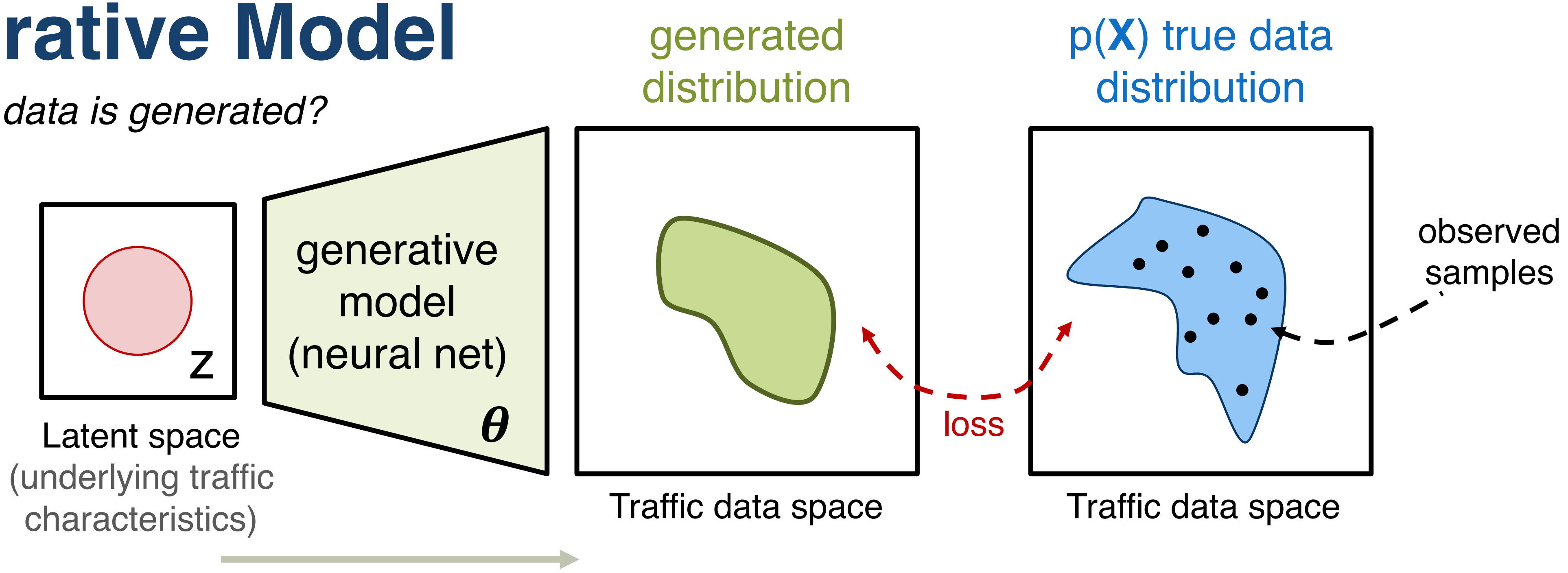
Solution:

- Generative model
- Bayesian inference to learn the data distribution and infer the missing values

→ multidimensional unsupervised online imputation method

Generative Model

How traffic data is generated?



X: traffic data (observed)
z: random latent variable
$\boldsymbol{\theta}$: model parameters

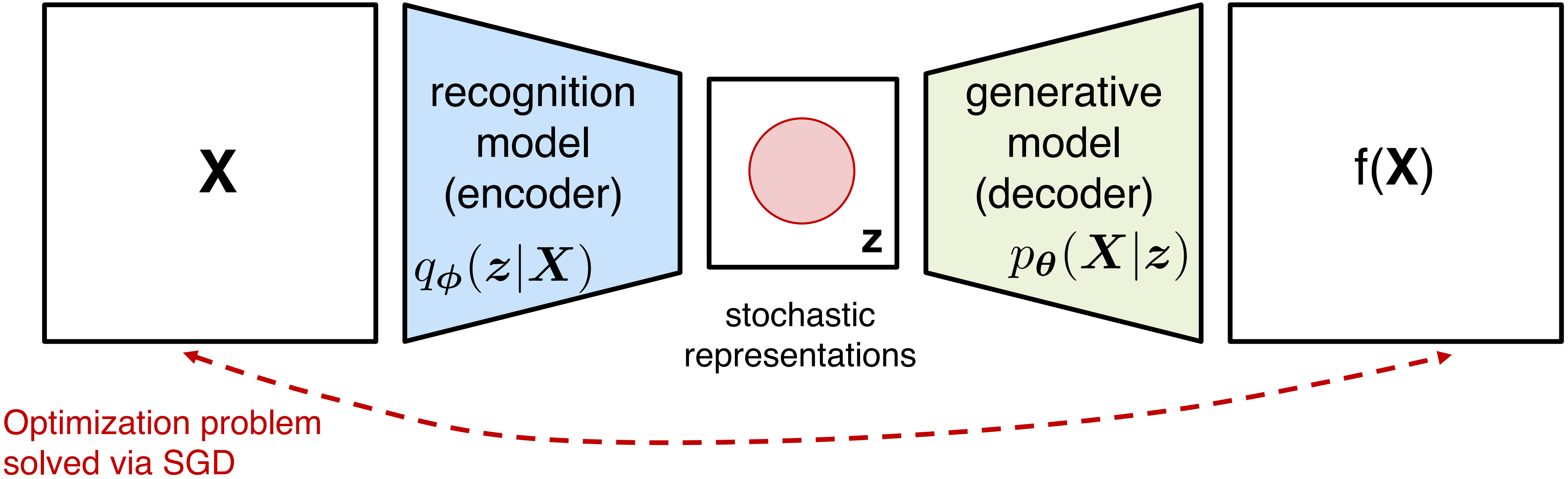
Maximum likelihood problem:

$$p_\theta(\boldsymbol{X}) = \int p_\theta(\boldsymbol{X}, z)dz = \int p_\theta(z)p_\theta(\boldsymbol{X}|z)dz$$

Intractable!

Variational Autoencoder (VAE)

[Kingma and Welling, 2014; Rezende et al., 2014]



$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$$

Reconstruction error

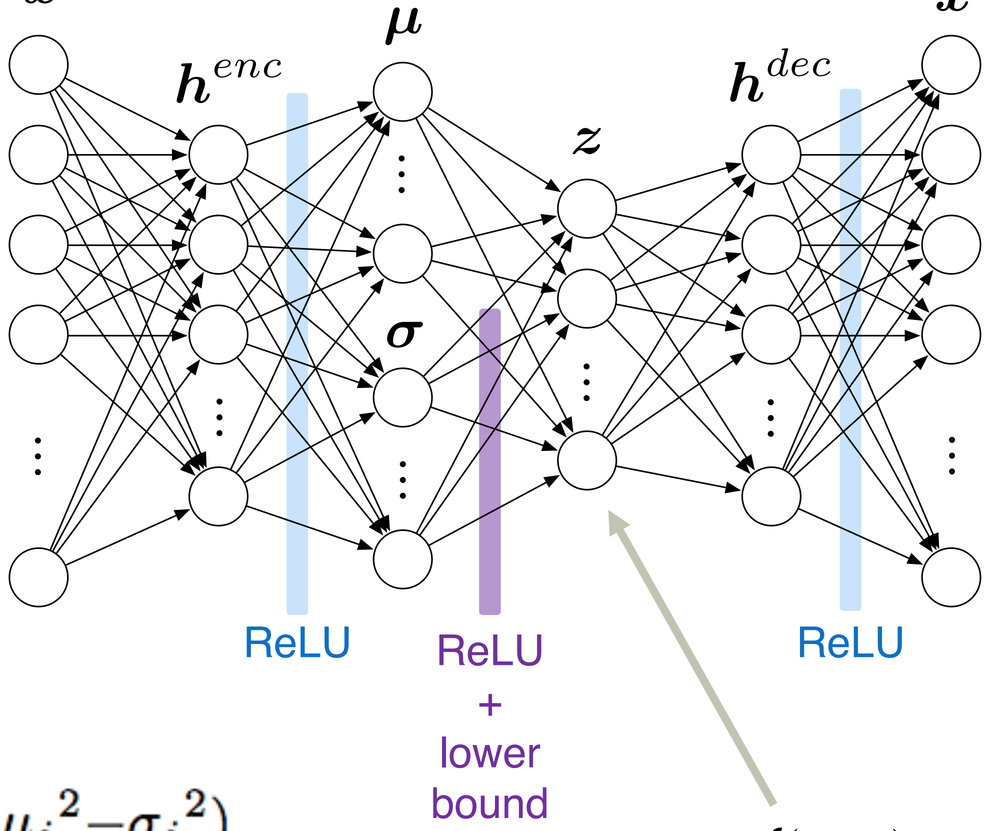
$$- D_{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \,\|\, p_{\boldsymbol{\theta}}(\boldsymbol{z}))$$

Regularizer:
approximate q to the true posterior p(**z**|**X**)

$$\leq \log p_{\boldsymbol{\theta}}(\boldsymbol{X})$$

Implementation

traffic sample: $\boldsymbol{x}$



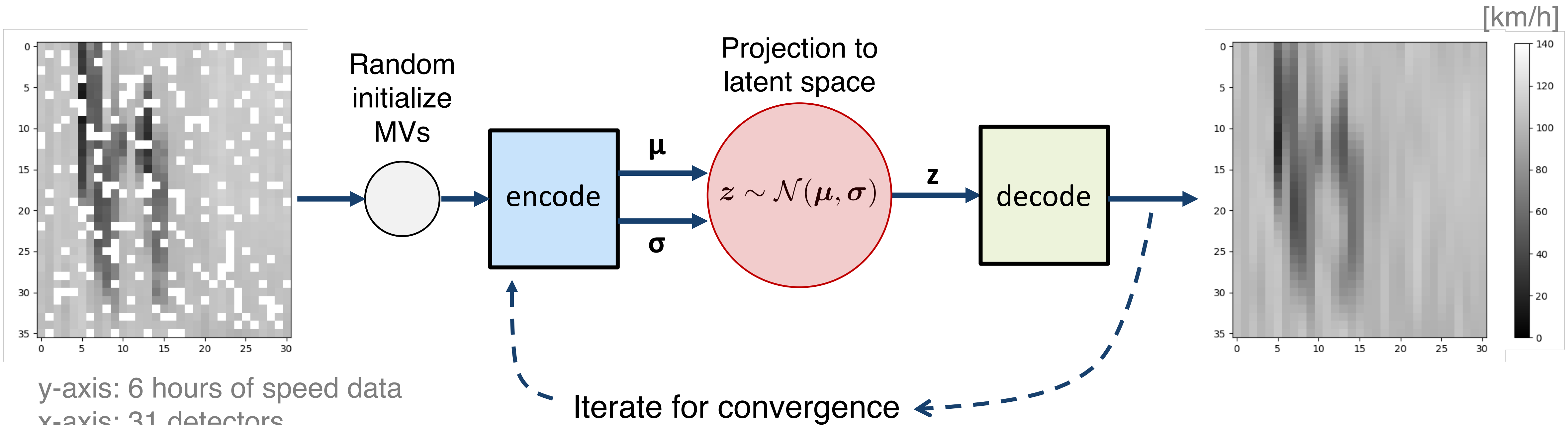
<u>Network assumptions</u>:
p(**z**) = Unit Gaussian
q(**z**|**X**) = Multivariate Gaussian
p(**X**|**z**) = Multivariate Gaussian

$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathbb{R}^J$
continuous latent space!

ϕ, ***θ***: *weights and biases?*

$$\arg\min \; \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; x) = \underbrace{||\boldsymbol{x} - \hat{\boldsymbol{x}}||^2}_{\text{MSE}} \underbrace{- \frac{1}{2}\sum_{j=1}^{J}(1 + \log {\sigma_j}^2 - {\mu_j}^2 - {\sigma_j}^2)}_{D_{KL}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) || \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}))}$$

Imputation procedure

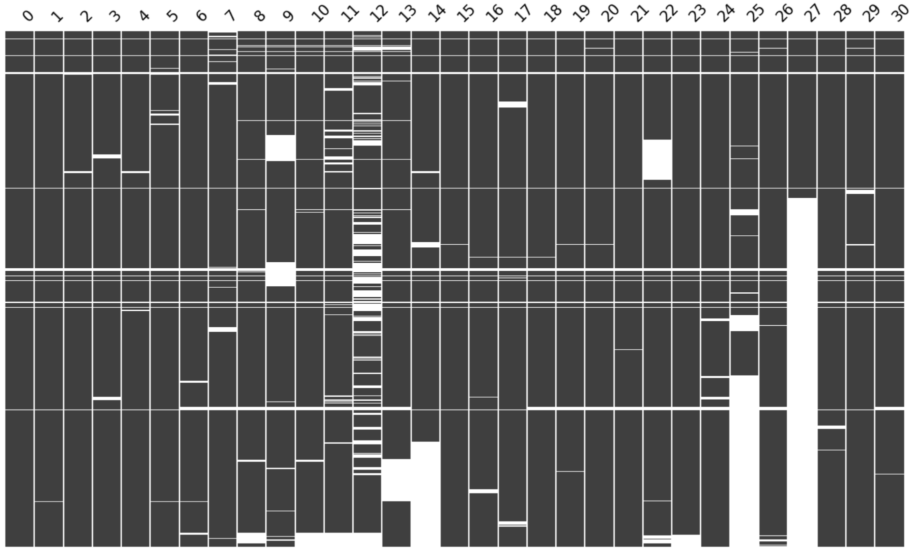


y-axis: 6 hours of speed data
x-axis: 31 detectors
white: MVs

(Markov chain that converges to the true marginal distribution of missing values given observed values [Rezende, 2014])

Real-world data set

Source: PeMS [http://pems.dot.ca.gov]
I-5 highway
31 sensors near San Diego
5-min samples from 2015 to 2017



NMAR (11.28% MVs)

Three data sets:

Original: PeMS imputed data [Chen, 2002]

Training: 2015 (105360 samples)

NMAR: samples with quality < 75% removed

MCAR-%: random 10, 20 and 40% removed

Testing: 2016 (105072 samples)

Experiment

Last 3 hours of traffic speed with missing values

$\tilde{\boldsymbol{x}} \in \mathbb{R}^{1116}$

Imputation layer

PCA [Li, 2014]

AE

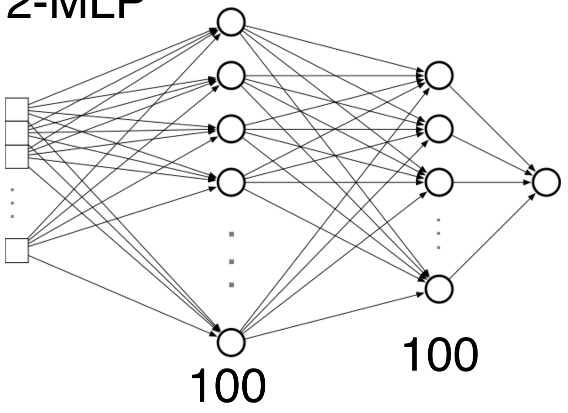
VAE

Reconstructed data

$\hat{\boldsymbol{x}}$

Regression layer (RL)

2-MLP



Future 1 hour traffic speed of sensor 15

$\boldsymbol{y}$

vs.

Original: $\boldsymbol{x}$

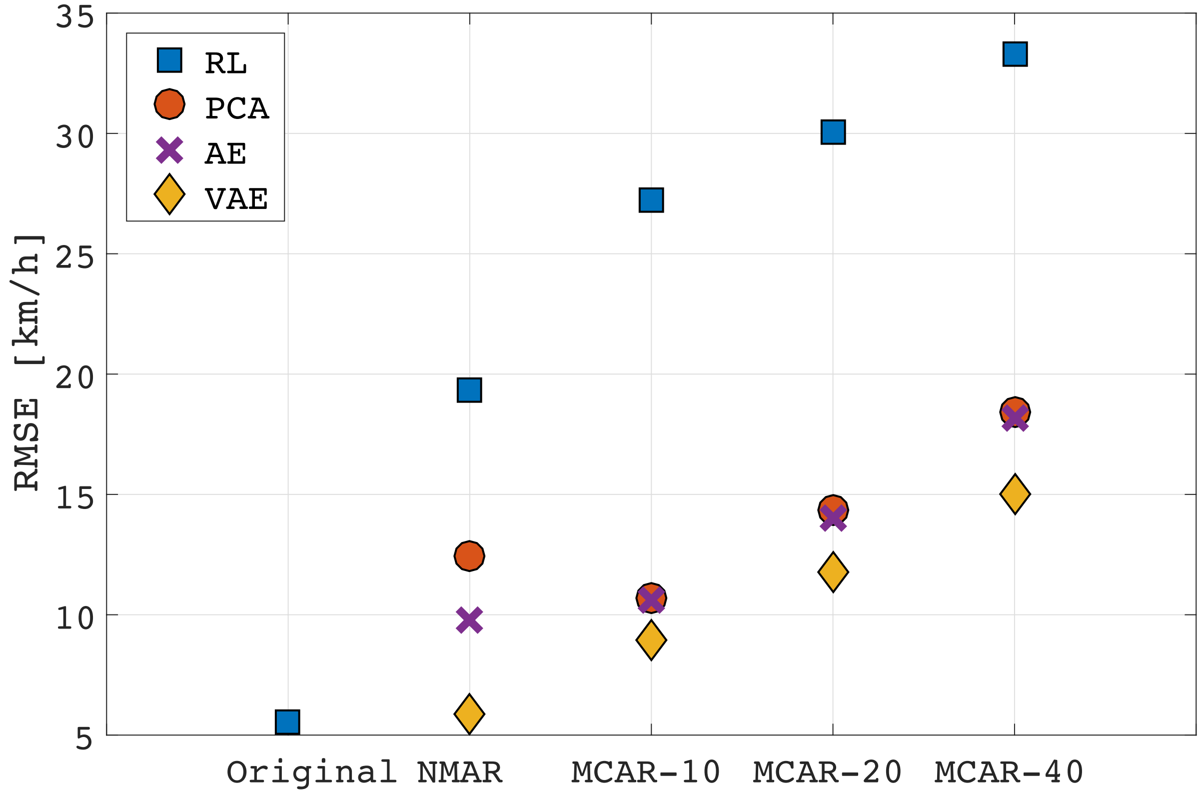
Corrupted: $\tilde{\boldsymbol{x}}$

$\boldsymbol{y}_{benchmark}$

Results

Impact on traffic forecast:

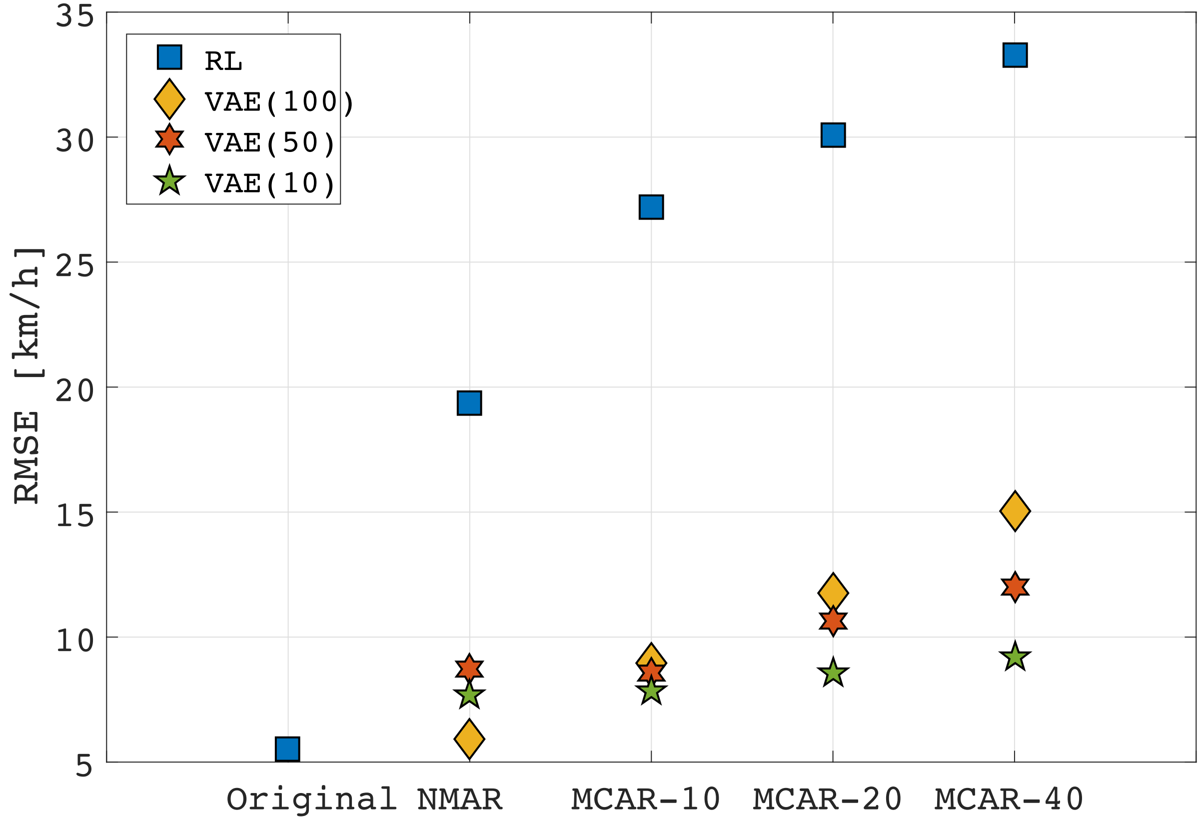
- RMSE improvement of **70%, 53%** and **40%** over RL, PCA and AE on **NMAR** data.
- RMSE improvement of **55%, 19%** and **17%** over RL, PCA and AE on **MCAR–40**.
- VAE performed better on NMAR (11.28% MVs) rather than MCAR–10



Results

Impact of code dimension:

- With a reduced code space dimension the accuracy remains similar despite increasing the MCAR proportion
- No significant results on NMAR data



Conclusion

- Multidimensional online unsupervised imputation method
- VAE can model traffic data and extract useful features
- Increases performance of traffic forecasting systems
- Improvements are greater on NMAR data which are mainly found on real-world data sets
- Also, useful for transportation modelers (future work):
    - Interpretability of the latent space (meaningful representations)
    - Outlier detection (anomalous traffic)
    - Dimension reduction (data compression)
    - Generative model with continuous latent space (road traffic network exploration)

QUESTIONS?

guillem.boquet@uab.cat

http://win.uab.cat