# Time Series Forecasting using RNNs

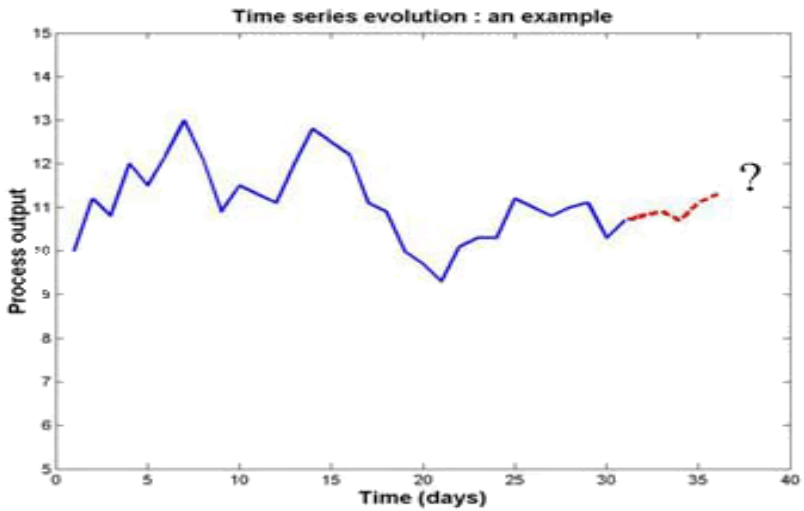
Y. Cinar, E. Gaussier, P. Goswami, H. Mirisaee
(M. Kaznacheeva for the presentation)

AMA, LIG, Univ. Grenoble Alpes

17 March 2017

# Table of Contents

# Time Series Prediction (1)



# Time Series Prediction (2)

- Real Time Series
  - Multivariate, multi-scale
  - May (or may not) contain several periods (seasonality, underlying patterns)
  - May (or may not) contain missing values (random and gaps)
- State-of-the-art methods
  - Ensemble methods: Random Forests, Gradient Boosted Trees (fixed length input)
  - Sequence-to-sequence RNNs (variable length input, machine translation, image captioning)
- Open questions
  1. Can RNN model periods in time series?
  2. Are they robust to missing values?

# Introduction to bidirectional RNNs with attention mechanism (1)

## Encoder

BiRNN: forward and backward RNNs for an input sequence **x**

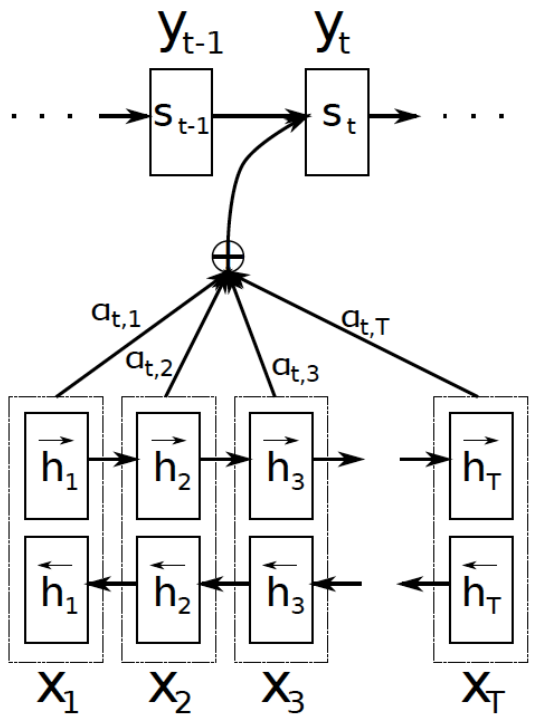
$$(x_1, ..., x_{T_x}) \Rightarrow (\overrightarrow{h}_1, ..., \overrightarrow{h}_T)$$

$$(x_{T_x}, ..., x_1) \Rightarrow (\overleftarrow{h}_1, ..., \overleftarrow{h}_T)$$

Final hidden state for each input $x_j$:

$$h_j = \begin{pmatrix} \overrightarrow{h}_j^\top \\ \overleftarrow{h}_j^\top \end{pmatrix}$$

In this way, the hidden state $h_j$ contains the summaries of both the preceding and following values



# Introduction to bidirectional RNNs with attention mechanism (2)

## Decoder

The probability is conditioned on a distinct context vector $c_i$ for each target word $y_i$:

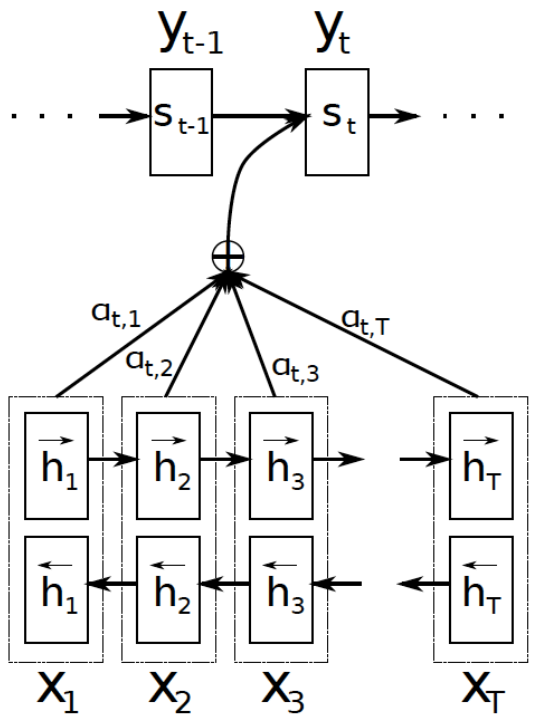
$$\mathbf{s}_i = g(y_{i-1}, \mathbf{s}_{i-1}, \mathbf{c}_i)$$

Context vector $c_i$ depends on all hidden states $(h_1, ..., h_T)$:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}$$

$$e_{ij} = \mathbf{v}_a^T tanh(\mathbf{W}_a \mathbf{s}_{i-1} + \mathbf{U}_a \mathbf{h}_j)$$



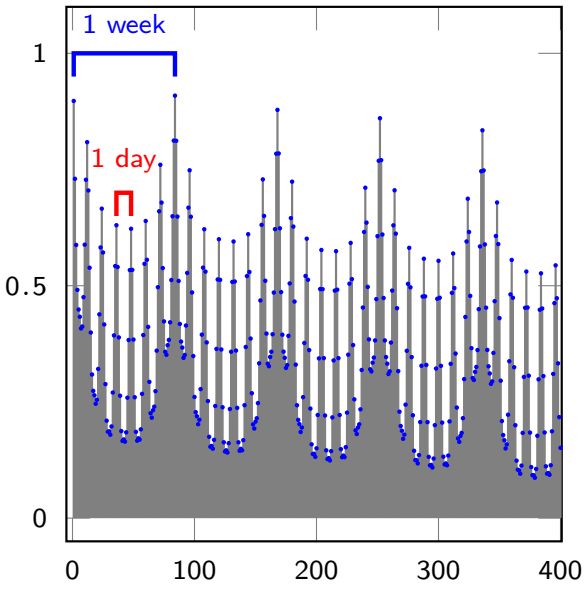
# Introduction to bidirectional RNNs with attention mechanism (3)

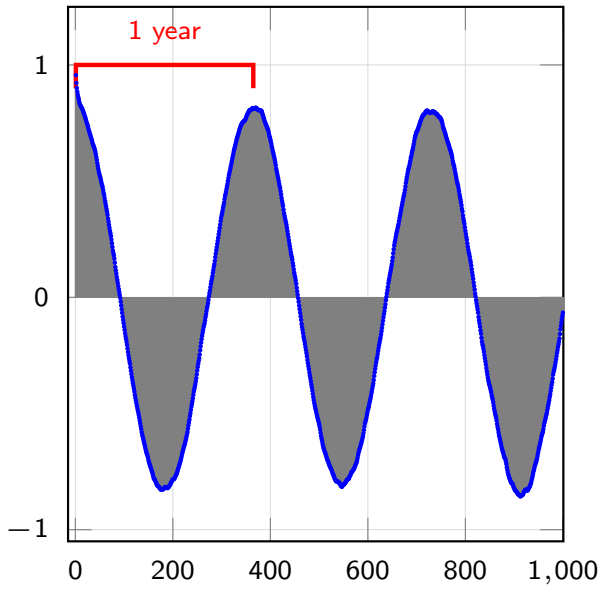
Eqs for decoder (LSTM)

$$
\begin{aligned}
\imath_i &= \sigma(\mathbf{W}_\imath y_{i-1} + \mathbf{U}_\imath s_{i-1} + \mathbf{C}'_\imath \mathbf{c}'_{i-1} + \mathbf{C}_\imath \mathbf{c}_i) \\
f_i &= \sigma(\mathbf{W}_f y_{i-1} + \mathbf{U}_f \mathbf{s}_{i-1} + \mathbf{C}'_f \mathbf{c}'_{i-1} + \mathbf{C}_f \mathbf{c}_i) \\
\mathbf{c}'_i &= f_t c'_{i-1} + i_i \tanh(\mathbf{W}_{c'} y_{i-1} + \mathbf{U}_{c'} \mathbf{s}_{i-1} + \mathbf{C}_{c'} \mathbf{c}_i) \\
\mathbf{o}_i &= \sigma(\mathbf{W}_o y_{i-1} + \mathbf{U}_o \mathbf{s}_{i-1} + \mathbf{C}'_o \mathbf{c}'_{i-1} + \mathbf{C}_o \mathbf{c}_i) \\
\mathbf{s}_i &= \mathbf{o}_i \odot \tanh(\mathbf{c}'_i)
\end{aligned}
$$

# RNN and periodicity (1)

Periodic time series: PSE, PW



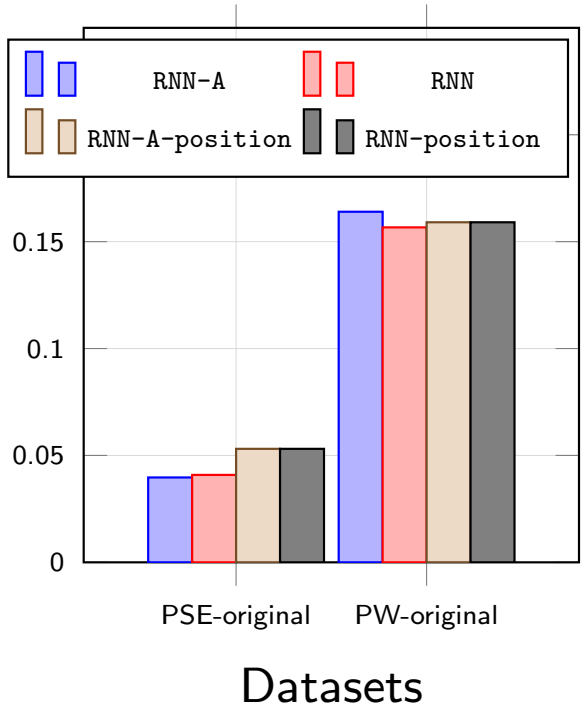




Figure 1: Autocorrelation for PSE and PW. A time span of 5 weeks is considered for PSE (left), and of 1.5 years for PW (middle). Comparison of RNNs with and without attention, and with and without time stamp (or position) information (right)

# RNNs and periodicity (2)

Attention weights as an (indirect) indication of the capacity to capture periods

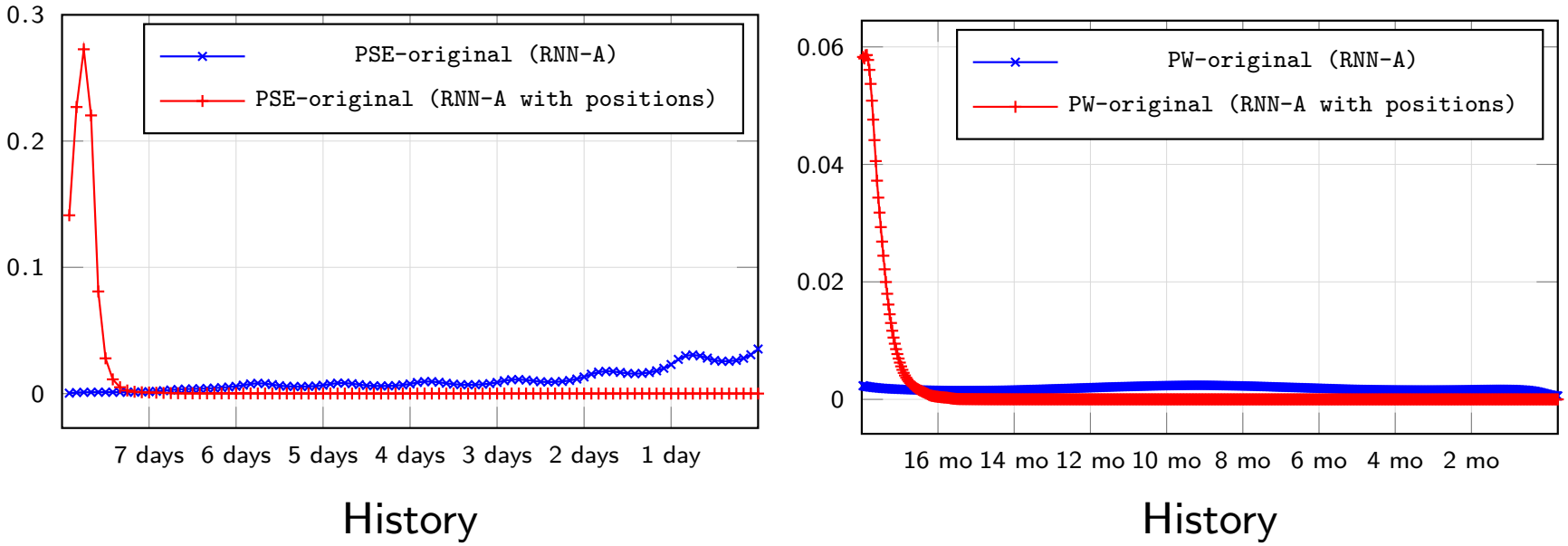
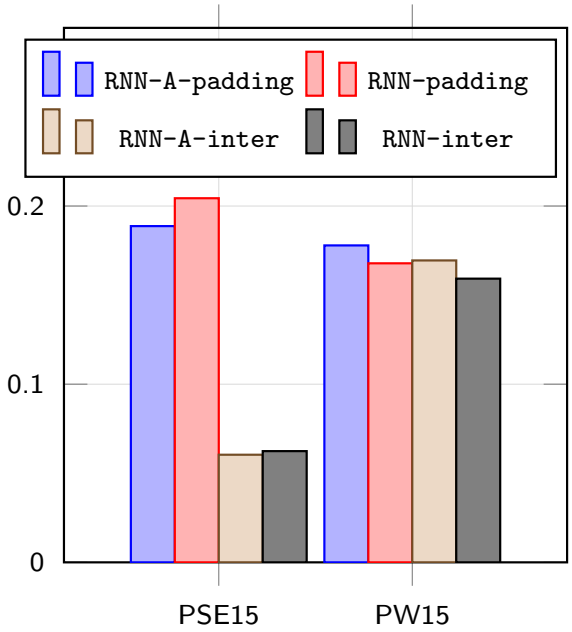



Figure 2: Weights of the attention mechanism for the RNNs without and with time stamp (or position) information on both PSE (left) and PW (right). The weights are averaged over all test examples

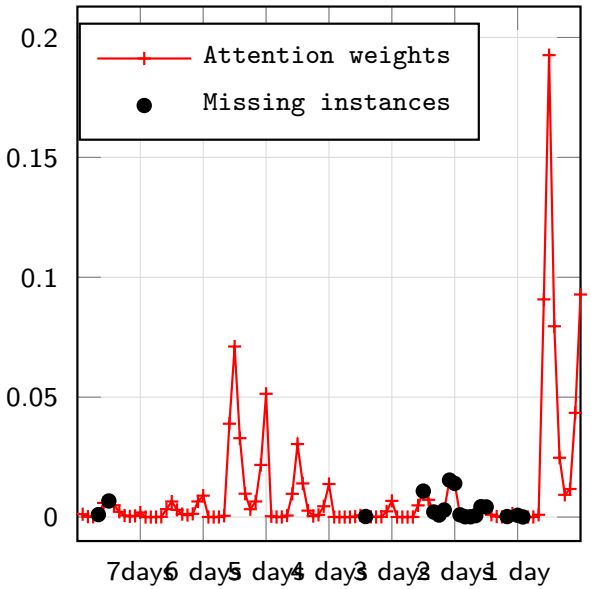
# RNNs and missing values (1)

- Usually possible to identify missing values (if sampling time does not change much)
- Two techniques:
  - interpolation (general as operates on values; linear, spline, kernel-based Fourier transform)
  - padding (RNN specific; repeats the last observed hidden state)

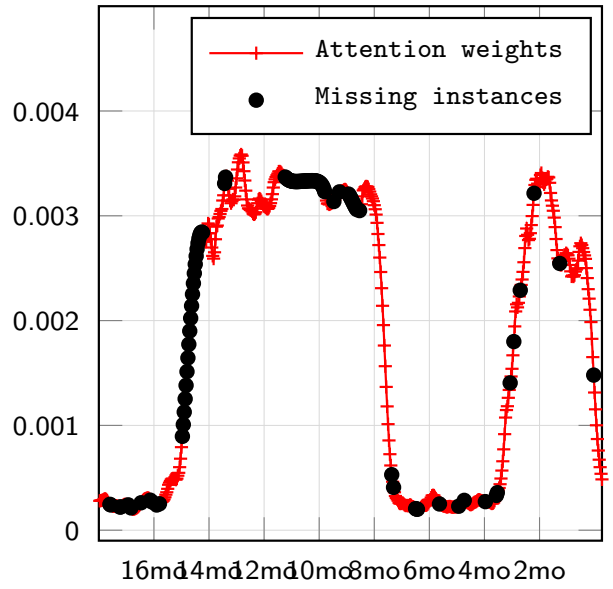
# RNNs and missing values (2)




Datasets




History

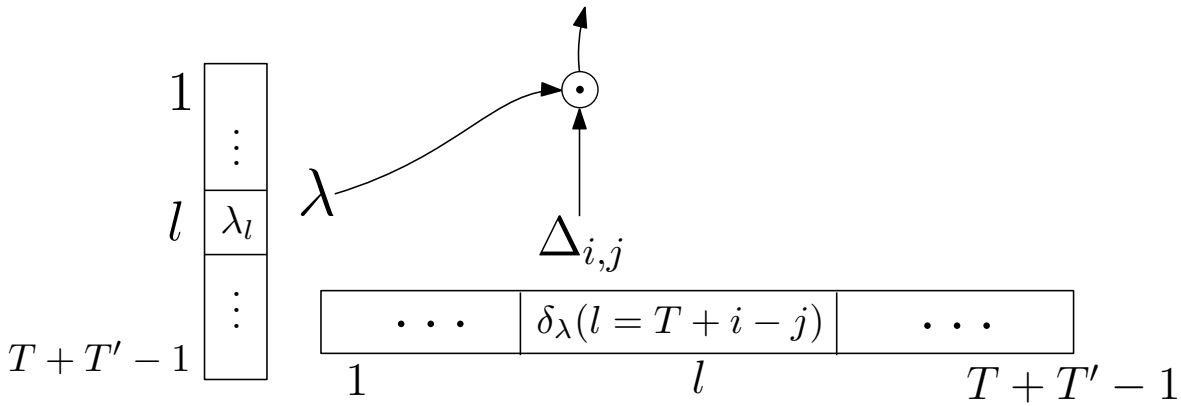



History

Figure 3: Comparison of padding and linear interpolation on PSE and PW with 15% missing values (left). Examples of attention weights on missing values for PSE and PW (right)

# Handling periods (1)

Explicitely model all relative positions and learn a weight to re-weigh the importance of given input according to relative position wrt output (RNN-$\tau$)




$$\Delta_{(i,j),l} = \delta_\lambda(l = T+i-j) = \begin{cases} 1 & \text{if } l = T + i - j \\ 0 & \text{otherwise} \end{cases}$$

Figure 4: Attention-$\lambda$

# Handling periods (2)

Same general attention mechanism as before, context vector $c_i$ depends on all hidden states $(h_1, ..., h_T)$:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T} \exp(e_{ik})}$$

With, this time:

$$e_{ij} = \underbrace{\mathbf{v}_a^T tanh(\mathbf{W}_a \mathbf{s}_{i-1} + \mathbf{U}_a \mathbf{h}_j) \times (\boldsymbol{\tau}^T \boldsymbol{\Delta}(i,j))}_{\beta_{ij}}$$

# Handling missing values

To handle the missing values, we consider two re-weighting schemes in the attention mechanism (resp. adapted to padding and interpolation):

- $e_{ij} = \beta_{ij} \times f_1(\mu, j)$

with

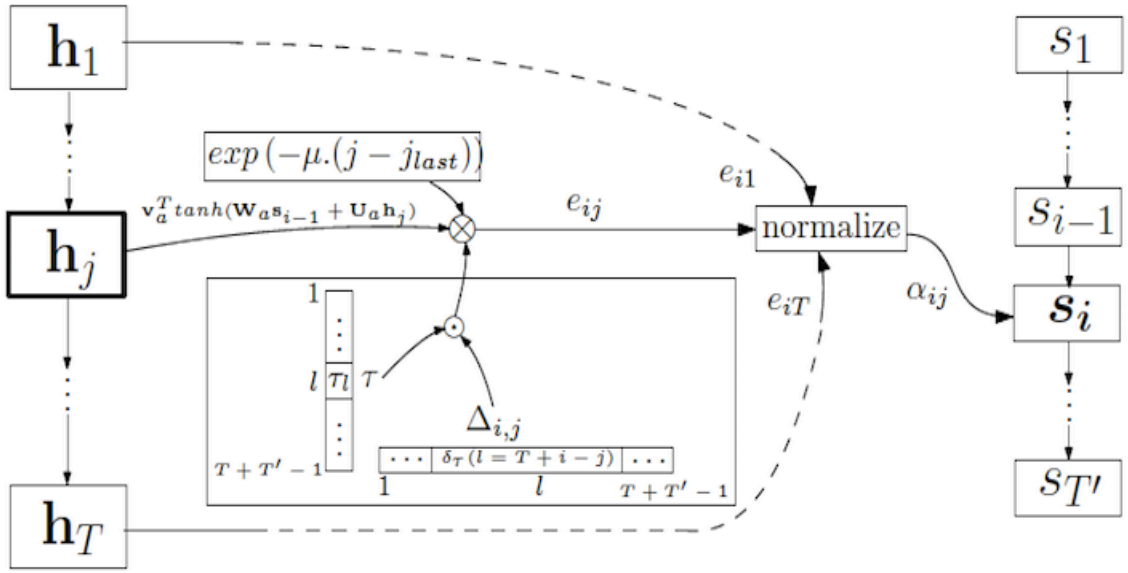
$$f_1(\mu, j) = \begin{cases} \exp(-\mu(j - j_{last})) & \text{if } j \text{ is missing} \\ 1 & \text{otherwise} \end{cases}$$

$j_{last}$ denotes the last observed value before $j$,
$\mu$ - learned parameter of RNN

- $e_{ij} = \beta_{ij} \times (1 + \mu^\top \mathbf{Pos}(j, \theta^g))$

$\mathbf{Pos}(j, \theta^g)$ - 3-dimensional vector which is null if $j$ is not missing; otherwise, one of the coordinates is set to 1 according to position of $j$ in the first, second or third equal parts of the gap $g$.

# Graphical illustration of the extended model

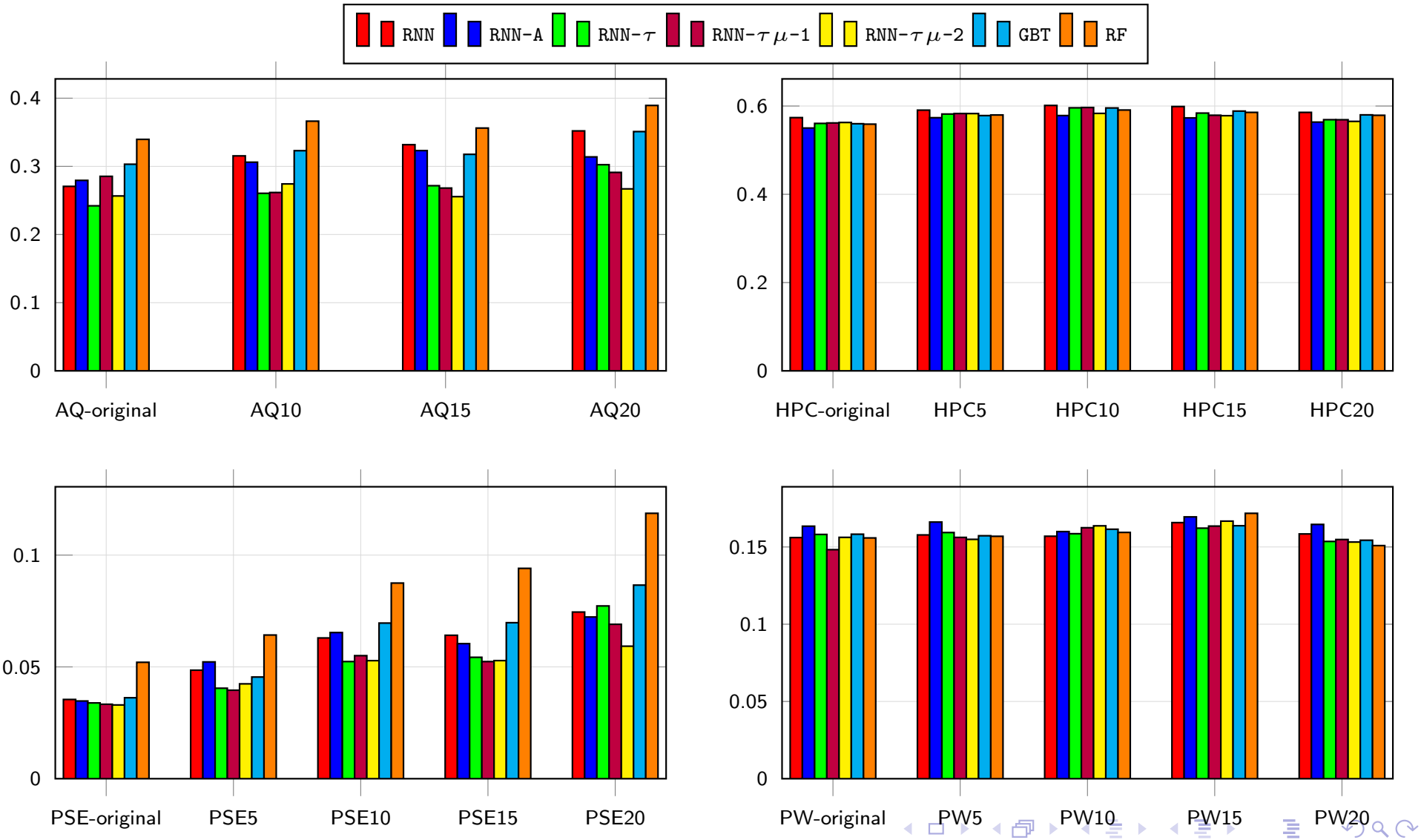


# Datasets

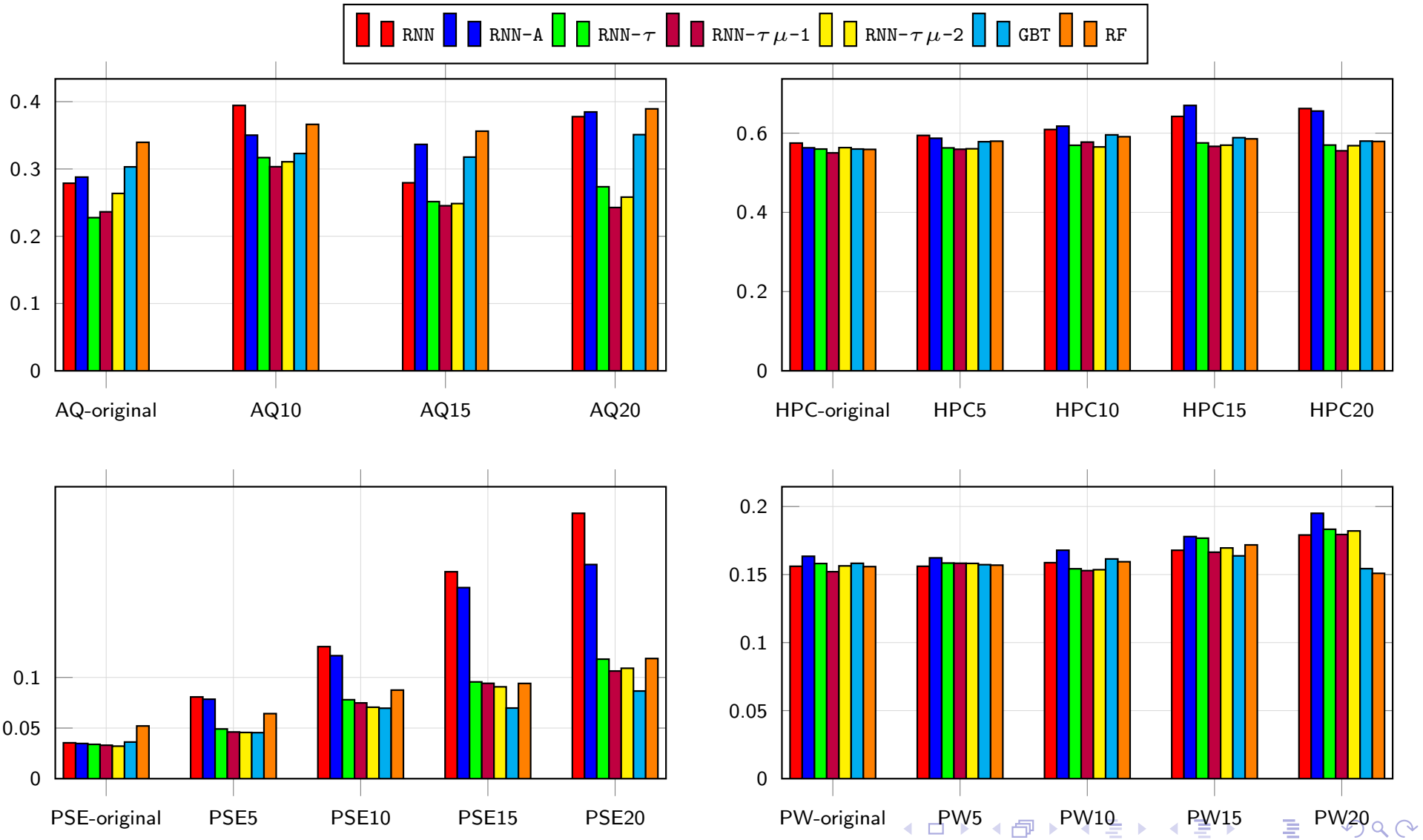
Table 1: Datasets.

| Name | #Inst. | Hist. size | For. hor. |
|---|---|---|---|
| Polish Elec. (PSE) | 46379 | 96 | 4 |
| Polish Weat. (PW) | 4595 | 548 | 7 |
| Air Quality (AQ) | 9471 | 96 | 6 |
| Hous. Power Cons. (HPC) | 17294 | 192 | 4 |

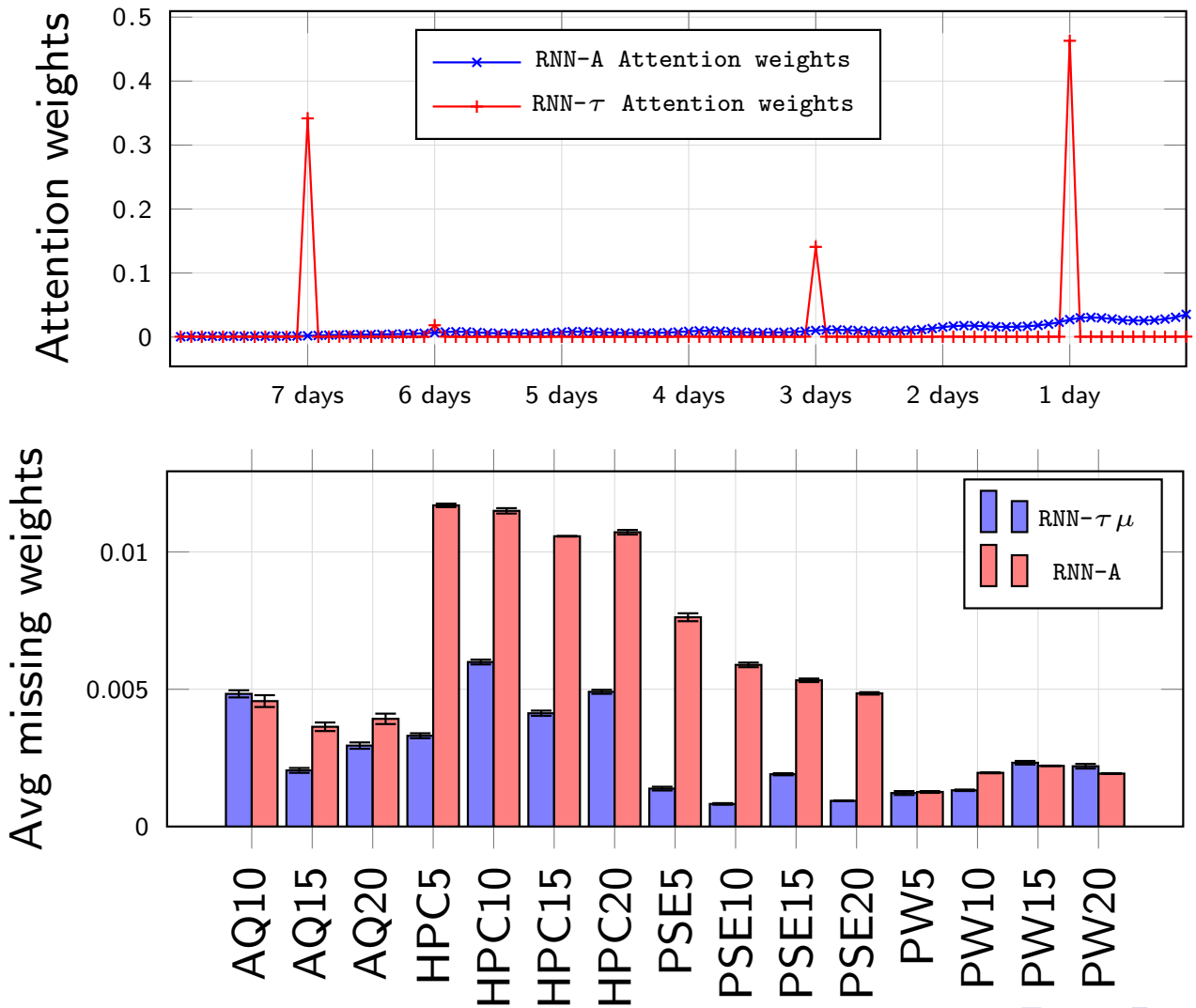
# MSE results on interpolated data



# MSE results on padded data



# Modeling periods and handling missing values



# Conclusion

- State-of-the-art RNNs (bidirectional LSTMs with attention mechanism) do not model well periods and are not robust to missing values (random and gaps)
- We have proposed two extensions:
  1. RNN-$\tau$: models periods through consideration of additional vector of relative positions
  2. RNN-$\mu$: avoids putting too much attention on missing values (for both padding and interpolation strategies)
- RNN)$\tau, \mu$ outperforms std RNNs, RFs and GBTs on datasets studied, especially when values are missing
- Similar results on multivariate extension

Thank you!

Questions?