STA2202- Time Series Analysis

Web Traffic Forecasting Using Singular Spectrum Analysis and Convolution Networks

An Exploratory Study

A Presentation By,

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Agenda

- ❑Defining the Problem
- ❑Data Description
- ❑Feature Engineering
- ❑Introduction of Methods
- ❑Modelling Approach
- ❑Model Identification
- ❑Results and Evaluation
- ❑Challenges
- ❑Conclusion
- ❑Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Defining the Problem

- Background:
  - Number of internet users has been growing worldwide, with total being 7.4 billion users, as of 2017, per report on ITU
  - With efficient data storage technologies, data has been become increasingly available at disaggregated levels
  - Forecasting at disaggregated levels, allow us to make more accurate predictions and capture local and global patterns in data
- Desired Outcome:

  A modelling approach to forecast web-traffic for different sites from different device and access agent
- Key Challenges:
  - High dimensions- a large set of pages, access agents and devices
  - Long range forecasting
  - Traditional time series models such as ARIMA will have a large set of parameters
  - Extreme events/ spikes in traffic due to external shocks

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

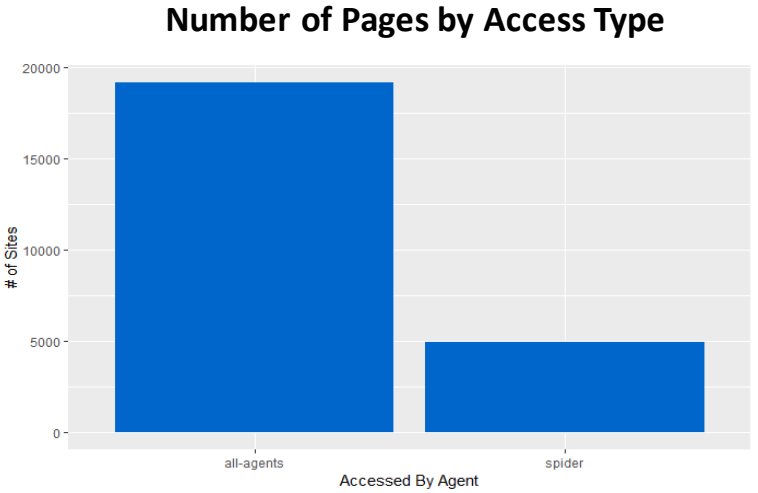
Agenda

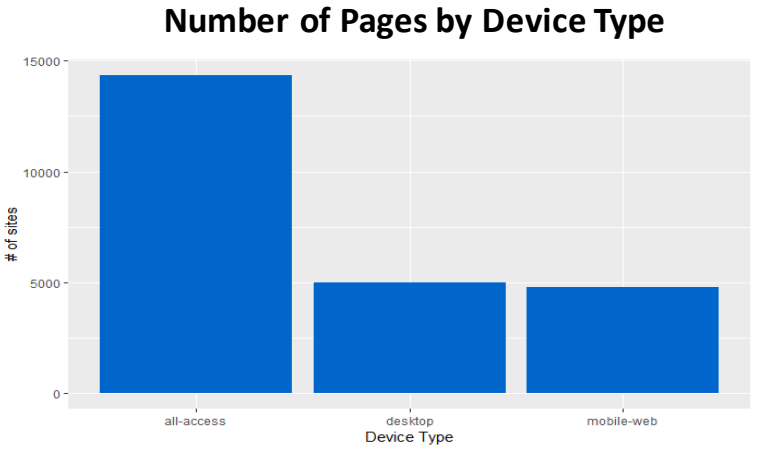
- ❑Defining the Problem
- ❑Data Description
- ❑Feature Engineering
- ❑Introduction of Methods
- ❑Modelling Approach
- ❑Model Identification
- ❑Results and Evaluation
- ❑Challenges
- ❑Conclusion
- ❑Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Data Description

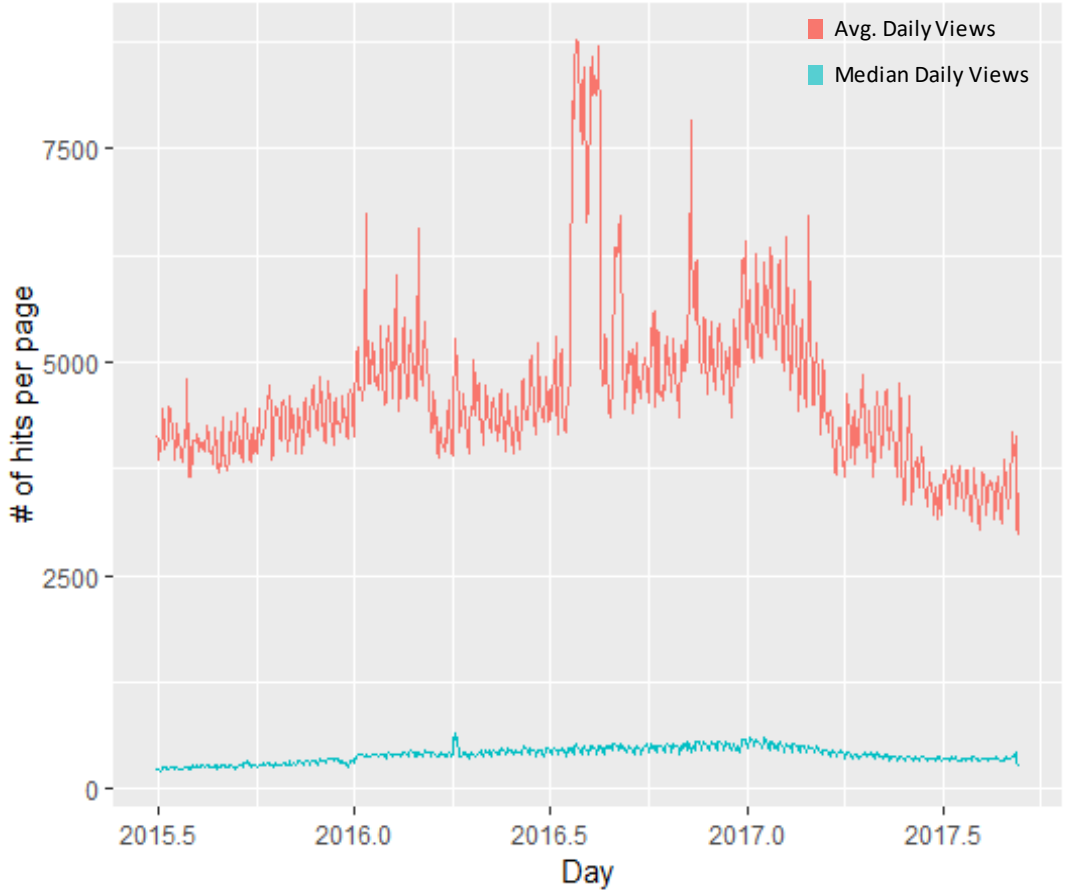
- Used a Kaggle dataset of 24K web-traffic series from Wikipedia English Project, from the duration of July 1st , 2015 to September 10th, 2017
- Level of the data: Webpage X Access Type X Device Type X Daily
- A NaN in data could be missing information or could be no visits on that day, no way to differentiate it.
  - Fill forward NaN and then replace remaining NaN with zeros- assuming there were no visits that day
- The data differentiates traffic from spider and all other agents- spider data is difficult to model, with occasional spikes
- The data differentiates traffic from Desktop, Mobile-Web from other devices



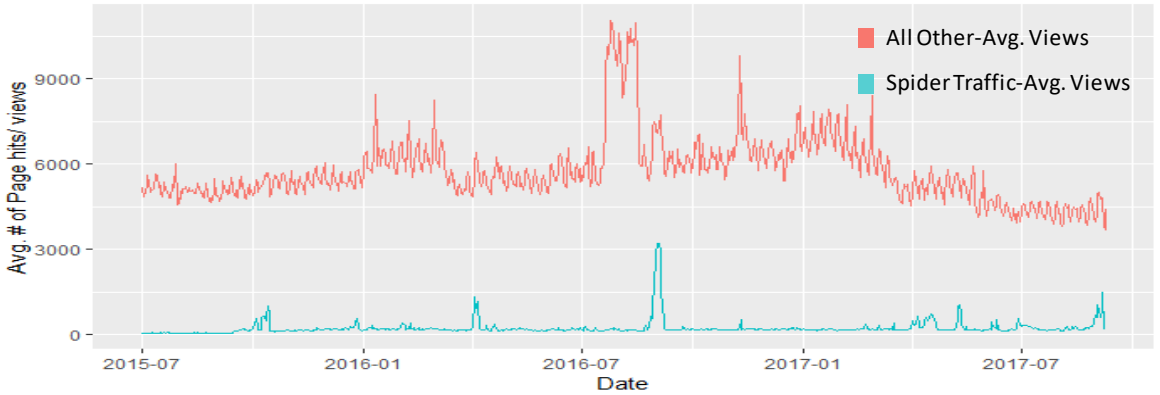


Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

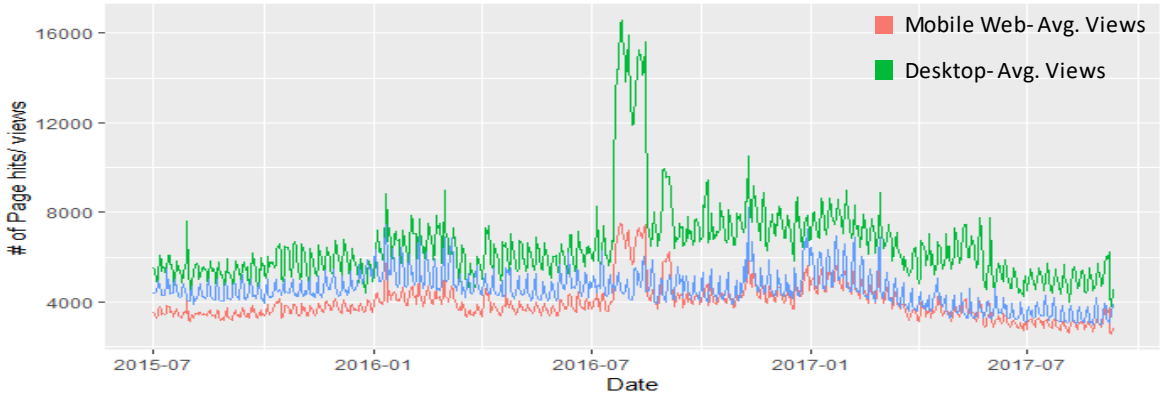
Data Description



Most pages are low noise traffic as can be seen from the median web views per days



Spider traffic experience occasional shocks, while other pages have clear periodicity in views



Desktop traffic is high volume as compared to mobile and other devices

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Agenda

- ❑Defining the Problem
- ❑Data Description
- ❑Feature Engineering
- ❑Introduction of Methods
- ❑Modelling Approach
- ❑Model Identification
- ❑Results and Evaluation
- ❑Challenges
- ❑Conclusion
- ❑Future Work

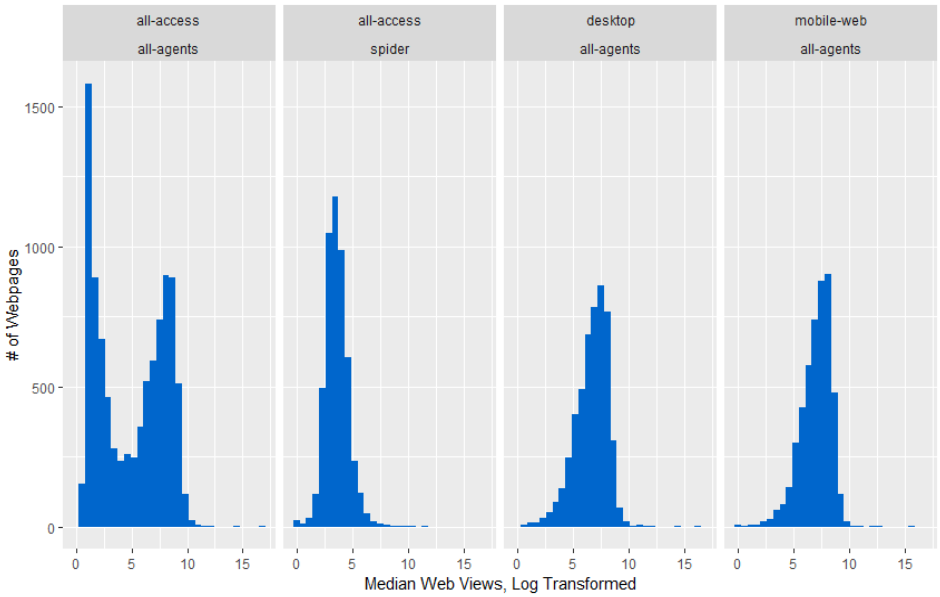
Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Feature Engineering

- The daily traffic views are Log-Transformed to remove effect of scales in multivariate data
- Most of the series log-transformed is stationary. However, 30% is still non-stationary
- Developed a one-hot encode for each access type and device type to capture inherent traffic flow from each source - For convolution networks
- The median traffic flow into each page, is built as a feature to differentiate high volume from low volume pages - For convolution networks
- A separate series for the overall traffic for Desktop, Mobile-Web, Spider is incorporated into the system, to capture the traffic inflow from different sources - For SSA

Based Augmented Dicky Fuller Test, for Log-Transformed Univariate series

Stationary	Non-Stationary
16,707 (70%)	7,245 (30%)



There are a lot of low noise traffic and few high noise traffic

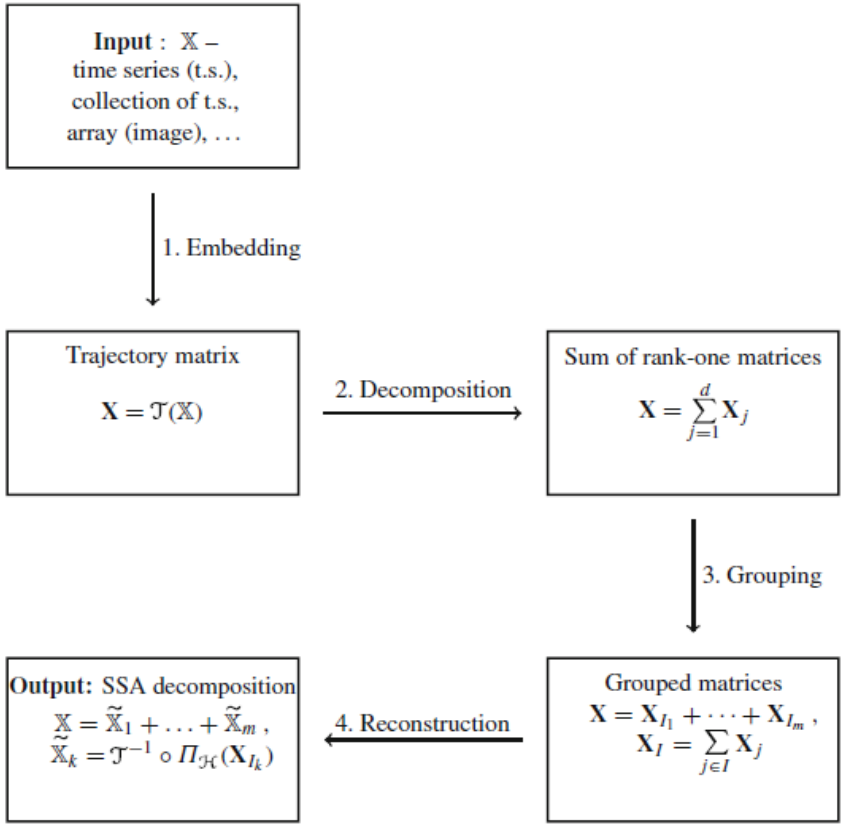
Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Agenda

- ❑ Defining the Problem
- ❑ Data Description
- ❑ Feature Engineering
- ❑ Introduction of Methods
- ❑ Modelling Approach
- ❑ Model Identification
- ❑ Results and Evaluation
- ❑ Challenges
- ❑ Conclusion
- ❑ Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Introduction of Methods: MSSA



1. **Embedding:** Construction of Trajectory Matrix

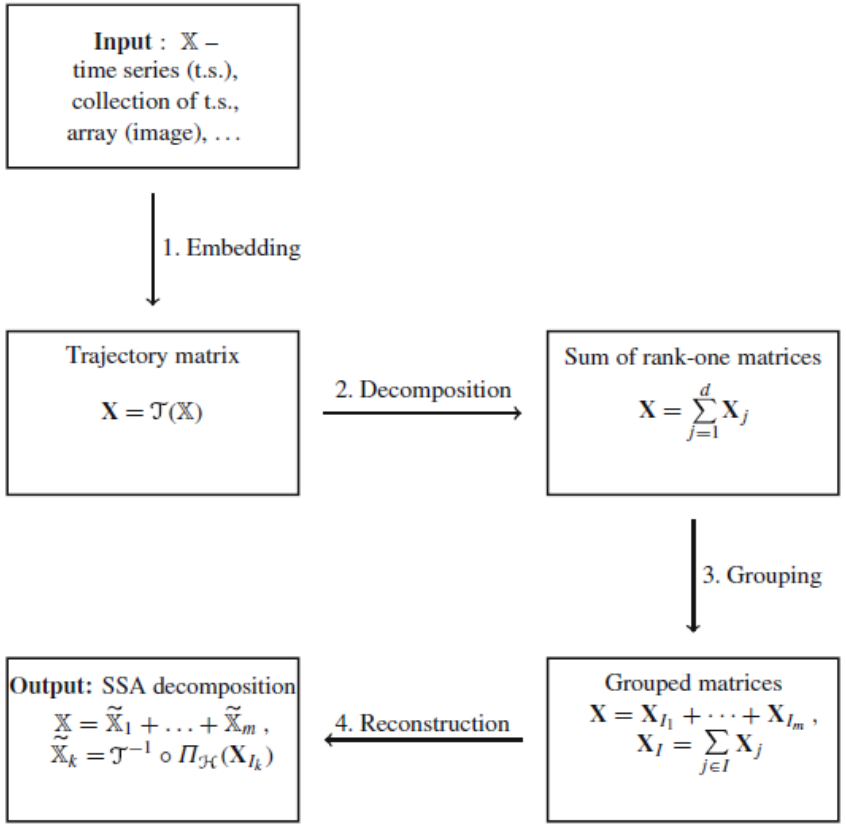
$$\tau(X) = \left[\boldsymbol{X^1} \ldots . \boldsymbol{X^s}\right],$$

$$where\ \boldsymbol{X}^p = \left(\boldsymbol{X_1^p} \ldots \boldsymbol{X_K^p}\right)^T = \begin{pmatrix} x_1^p & \ldots & x_K^p \\ x_2^p & \ldots & x_{K+1}^P \\ x_L^p & \ldots & x_N^p \end{pmatrix}$$

L = Window length, N = Total number of observations, p = 1.. s1 < L < N and K = N − L + 1

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Introduction of Methods: MSSA



2. **Singular Valued Decomposition:**

$$\boldsymbol{X = X_1 + \cdots + X_d; X_i = \sqrt{\lambda_i} U_i V_i^T}$$

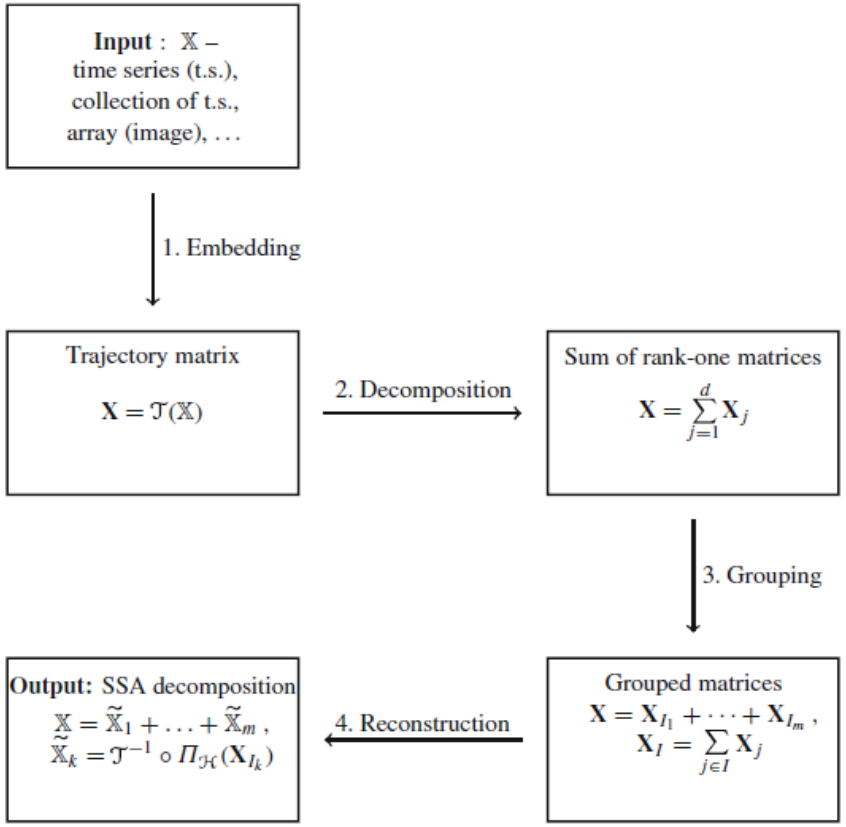
$$where\ U_i\ \forall i = 1..d\ denote\ eigen\ vectors\ V_i = \frac{X^T U_i}{\sqrt{\lambda_i}}$$

$$\lambda_i\ eigen\ value\ o\ f\ \forall i = 1..L\ be\ eigen\ values\ S = XX^T$$

$$d = \max\{i; \lambda_i > 0\}$$

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Introduction of Methods: MSSA

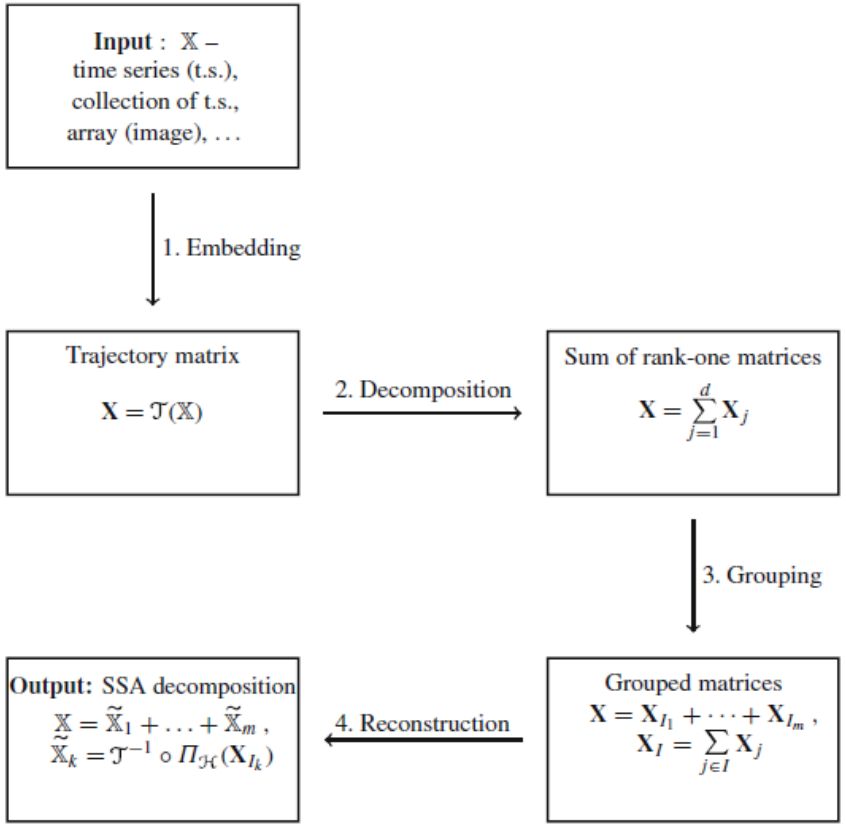


3. **Grouping:**

Partition the set of indices {1, . . . , d} into m disjoint subsets $I_1, \ldots, I_m$; $I = \{I_1, \ldots, I_p\}$
Then the resultant matrix X_I corresponding to the group I is defined as $X_I = X_{I_1} + \ldots + X_{I_p}$

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Introduction of Methods: MSSA



4. **Reconstruction:**

Perform diagonal averaging/ Hankelization to transform the reconstructed matrix $\boldsymbol{X_{I_m}}$ to the form of a Hankel matrix

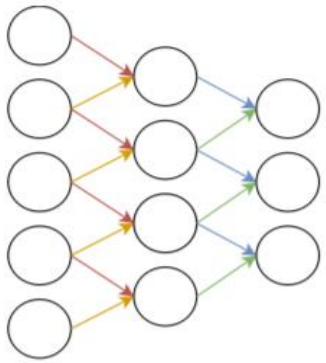
$$X = X_1 + \cdots X_m; X_m = \tau(X)^{-1} * \pi(X)$$

$\pi(x)$ is the orthogonal projector

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Introduction of Methods: Convolution Networks

- Commonly applied to image and speech classification problems
- **Key features:** local connectivity and shared parameter space
- Input convolved with a set of filters applied over all the input channels, to create the feature output map. Filter size controls the receptive field of the network
- Captures local patterns and extracting similar features across data using kernels/ filters which slide across the data in fixed widths to generate abstract features from the series
- Optimal weights are found through gradient descent technique.
- Run through a number of iterations, each with a forward (compute forecasts) and a backward pass (update weights)



Filter size:1X 2 | 2 layer CNN
Receptive field of each node consists of two input neurons from the previous layer and weights are shared across the layers

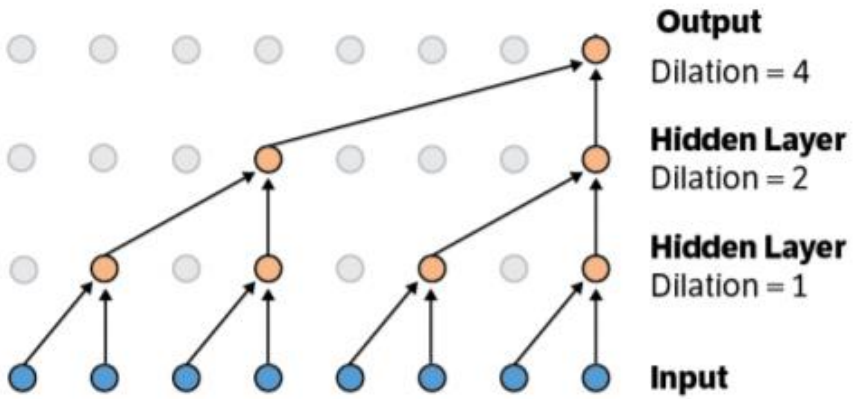
[Picture taken from paper: Conditional time series forecasting with convolutional neural networks]
[**ref:** https://arxiv.org/pdf/1703.04691.pdf]

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Introduction of Methods: Convolution Networks

- Core structure of a Wave-Net is causal padding and dilation
- The time series data is padded such that the prediction at time step, t doesn't depend on any of the future time steps, (t+1), (t+2) and so on
- The idea of Wave-Net is to use CNN as an AR(p) model of form,

$$\hat{x}(t+1) = \sum_{i=1}^{p} \alpha_i x_{t-i} + \epsilon(t); where\ \epsilon(t)\ is\ WN$$



[Picture taken from paper: Conditional time series forecasting with convolutional neural networks]
[**ref:** https://arxiv.org/pdf/1703.04691.pdf]

Agenda

- ❑Defining the Problem
- ❑Data Description
- ❑Feature Engineering
- ❑Introduction of Methods
- ❑Modelling Approach
- ❑Model Identification
- ❑Results and Evaluation
- ❑Challenges
- ❑Conclusion
- ❑Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Modelling Approach

Input Series

Multivariate Series of dimension, T X M

M: Number of Series
T: Time

Trend Extraction

Construct the Trajectory Matrix

Decomposition

Grouping

Reconstruction

Seasonality Extraction

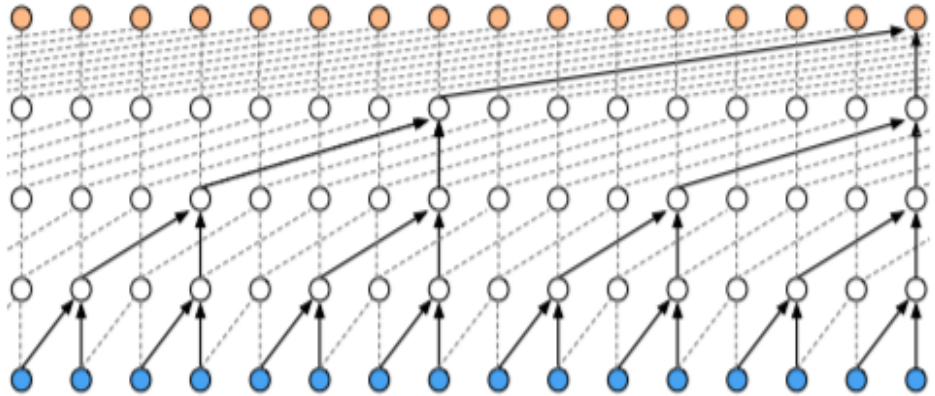
Construct the Trajectory Matrix

Decomposition

Grouping

Reconstruction

Residual Modelling



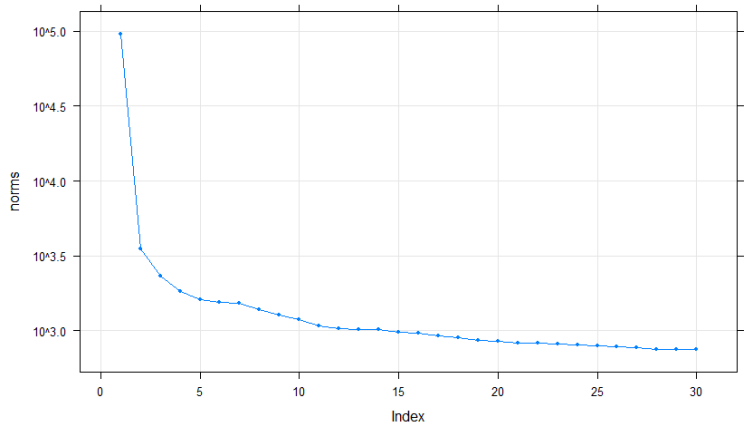
Agenda

- ❑ Defining the Problem
- ❑ Data Description
- ❑ Feature Engineering
- ❑ Introduction of Methods
- ❑ Modelling Approach
- ❑ Model Identification
- ❑ Results and Evaluation
- ❑ Challenges
- ❑ Conclusion
- ❑ Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

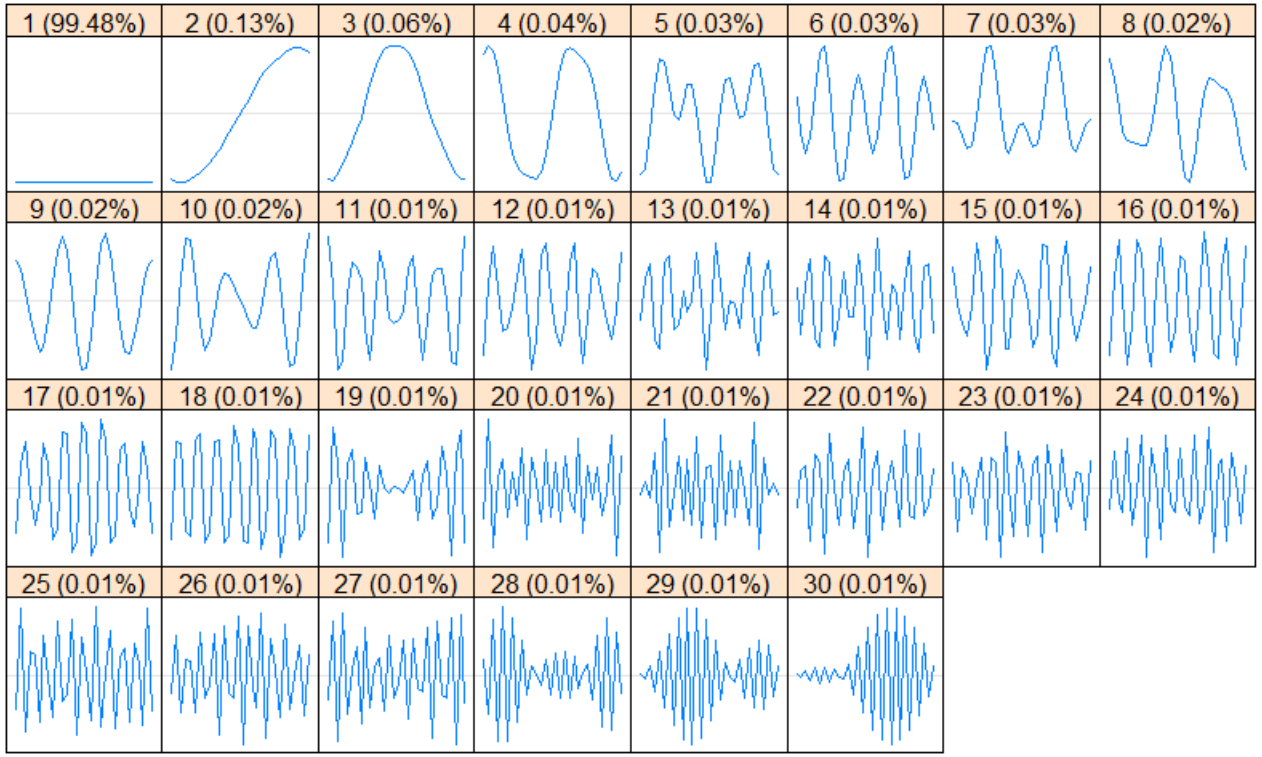
SSA-Trend: Model Identification

Scree Plot



Sharp drop in slope at 1st eigen vector, indicates it should be treated as a separated component in itself

Eigen Vectors

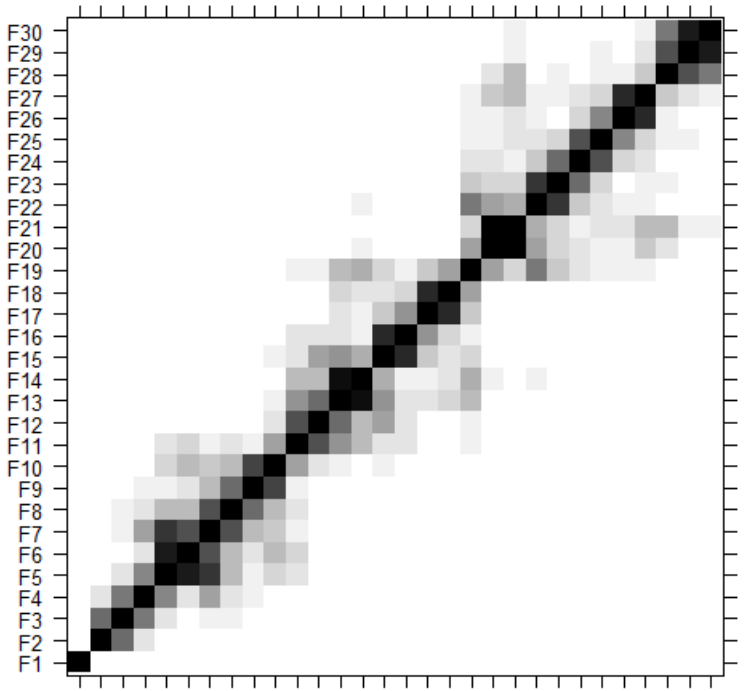


Looking at the Eigen Vector plot, it is seen that most of the component in train series is simple constant median, about 99%

Chose a small Window Length for this SSA, to primarily intend to extract Trend

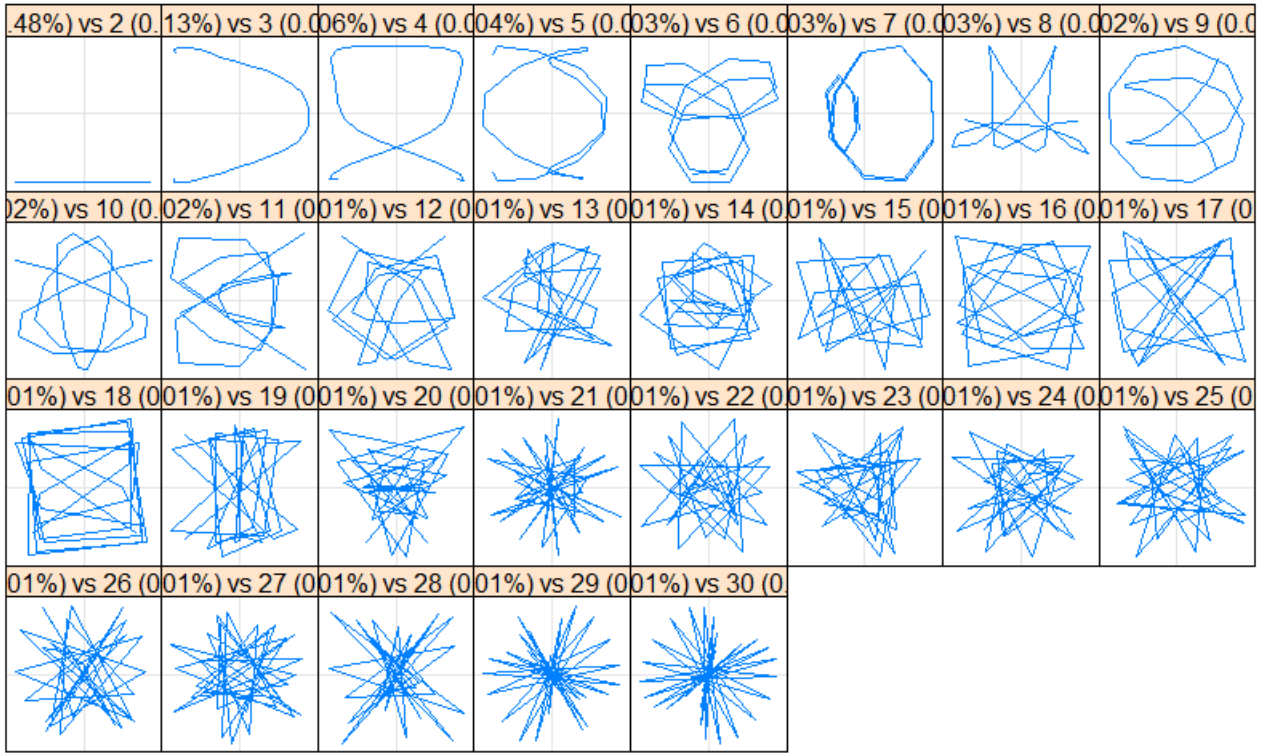
SSA-Trend: Model Identification

W- Correlation Matrix



As can be seen, the first component is well-separable from the rest

Paired Eigen Vectors

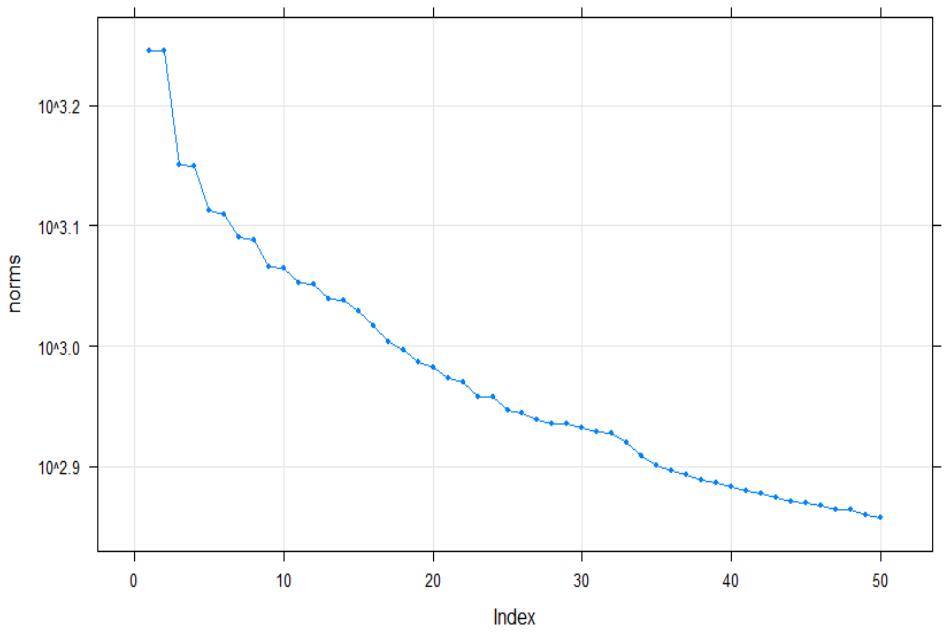


No pronounced seasonality components can be identified from decomposition, this could also be because of the components not being well-separable

Remove the trend components and feed remaining parts of series into another SSA

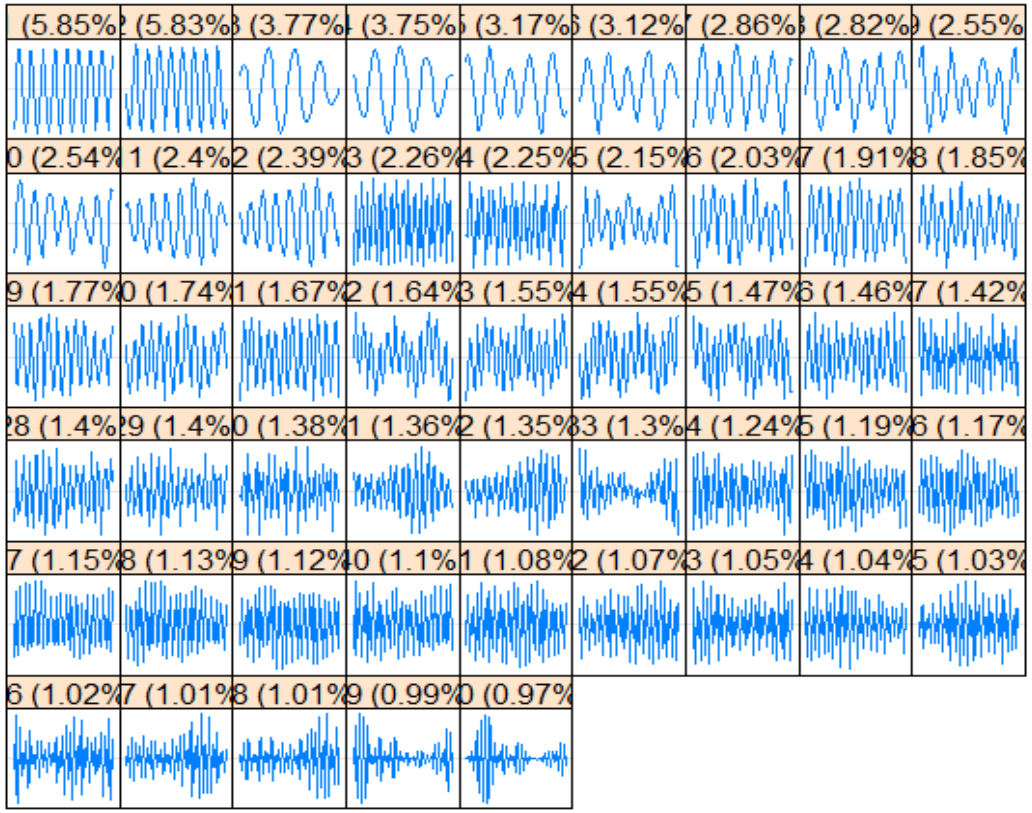
SSA-Seasonality: Model Identification

Scree Plot



Chose a larger Window Length for this SSA, to extract periodic components

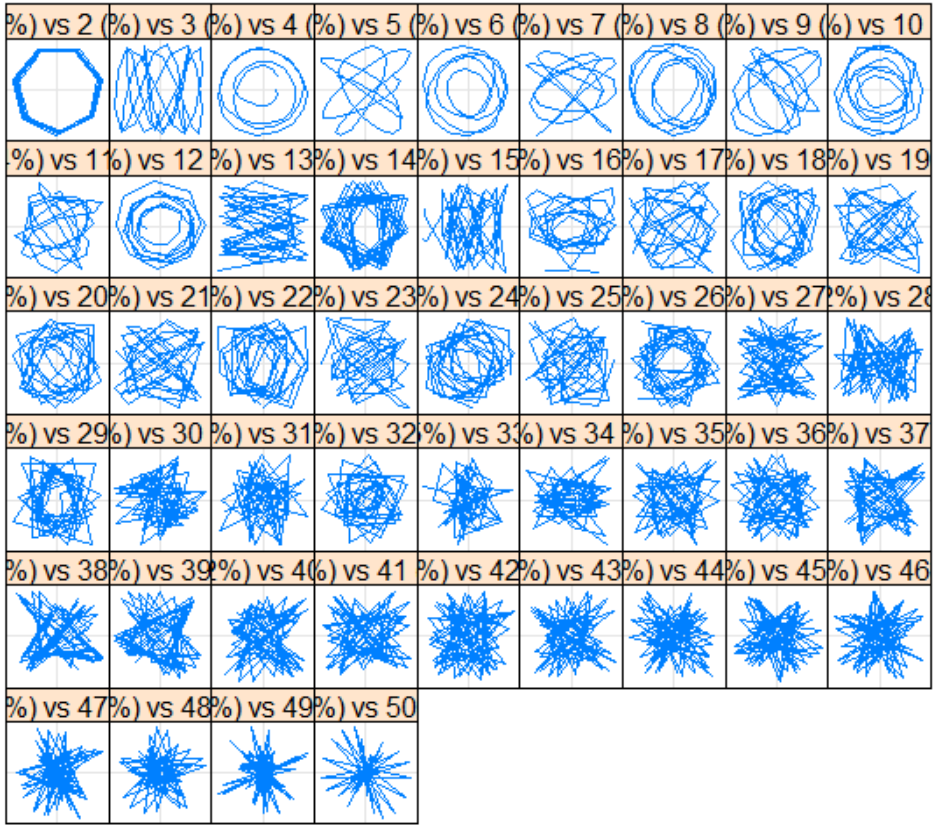
Eigen Vectors



1 and 2 component have similar norms and is distinctive from the rest, we treat them as one pair

SSA-Seasonality: Model Identification

Paired Eigen Vectors

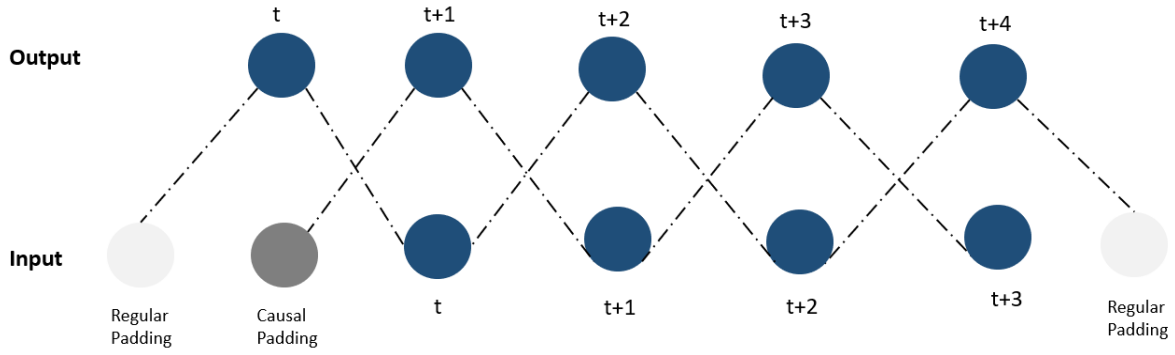



```
> parestimate(ssa_remaining, groups = list(1:2), method = "esprit")
   period      rate    |     Mod      Arg   |      Re        Im
    7.036  -0.000752 |  0.99925    0.89 |  0.62664   0.77835
   -7.036  -0.000752 |  0.99925   -0.89 |  0.62664  -0.77835
```

As is evident from the pair plots and the parameter estimate,1 and 2 component represent weekly seasonality

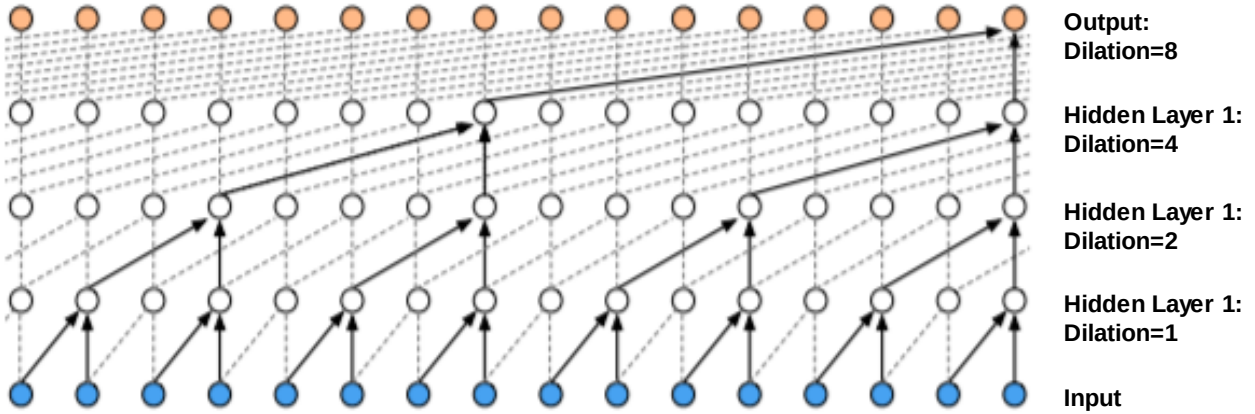
Residual Modelling using Convolution Networks

- The residuals from the Sequential SSA model are fed into a 1 D convolution network with 9 stacked layers, and dilation exponentially increasing at each layer
- With a Kernel of size, 7, capture correlations within series in 7 time steps
- To ensure, forecasting at time t, doesn't depend on t+1,t+2,… observation, a causal structure is used which pads the series data asymmetrically
- Forecasting of series x_{t+1} is conditioned on the actual web-views through $x_{1,2,\ldots t}$ and the web-views of the other pages $y_{1..t}^{p}\ \forall\ p = 1..s, \text{where } s \text{ is the number of webpages}$. This is achieved using a ReLU activation function



Residual Modelling using Convolution Networks

- Additionally, we incorporate features to identify inherent traffic variations from Desktop, Mobile and Spider-using one-hot encode features and median views as another feature
- The data is reshaped to a tensor of dimension, [Number of Pages * Day * Features]
- A 70-30 validation split is used, with a batch size of 256
- Early Stopping is used, if the validation loss doesn't decrease after 10 epoch. The model is trained for a maximum of 2,000 epochs
- A dropout layer is used to handle overfitting, with a dropout rate of 0.2
- An illustration of the network for 4 dilation layers is as follows,



[Picture taken from paper: WaveNet: A Generative Model for Audio]
[**ref:** https://arxiv.org/pdf/1609.03499.pdf]

Agenda

- ❑ Defining the Problem
- ❑ Data Description
- ❑ Feature Engineering
- ❑ Introduction of Methods
- ❑ Modelling Approach
- ❑ Model Identification
- ❑ Results and Evaluation
- ❑ Challenges
- ❑ Conclusion
- ❑ Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Results and Evaluation

- SMAPE (Symmetric MAPE) was used as validation metric.
  - Robust to outliers
- SMAPE= 2*(|Actual-Forecast|)/(|Actual|+|Forecast|)
- We trained on one year of data and forecast for the next 64 days, since we have two years of information from 2015-07-01 to 2017-09-10, we can then do a slide window cross validation and evaluate how our model performs at different points of time

Comparing three approaches, for latest validation period of 2017-05-17 to 2017-07-19,
Trained on 1 year of prior data 2016-05-16 to 2017-05-16

Modelling Approach	Forecast Horizon (in days)									
	0	7	14	21	28	35	42	49	56	63
CNN Only	52.90	53.28	52.90	53.16	53.62	53.55	53.58	53.73	53.71	53.65
Seq. SSA only	25.58	29.06	30.77	31.65	32.70	33.90	34.78	35.60	36.15	36.44
Seq. SSA + CNN	**25.36**	**28.84**	**30.48**	**31.33**	**32.38**	**33.55**	**34.42**	**35.22**	**35.78**	**36.12**

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Results and Evaluation

Performing side walk validation for 1 year prior train and 64 days ahead forecast across different points of time in history start from July 1st 2015, for a combined SSA+CNN approach

Validation Period	Avg. SMAPE
01-07-2016 - 02-09-2016	44.53
03-09-2016 - 05-11-2016	43.41
06-11-2016 - 08-01-2017	42.41
09-01-2017 - 13-03-2017	42.08
14-03-2017 - 16-05-2017	36.73
17-05-2017 - 19-07-2017	36.12

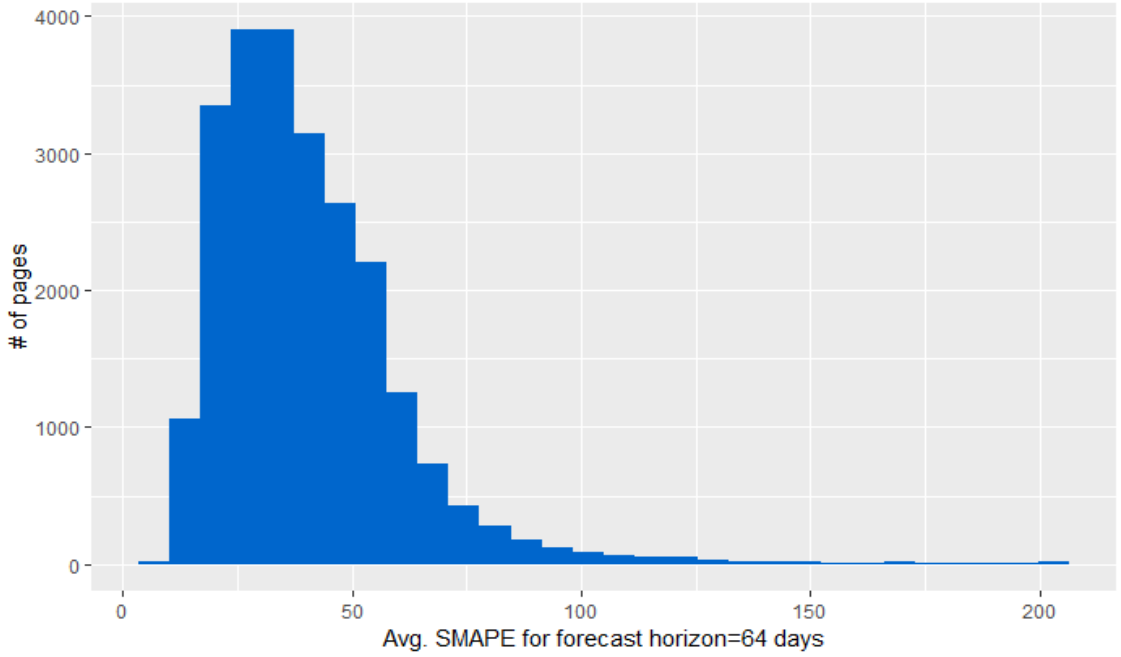
Model is improving over time in terms of forecast accuracy

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Results and Evaluation

Simple Univariate ARIMA Model

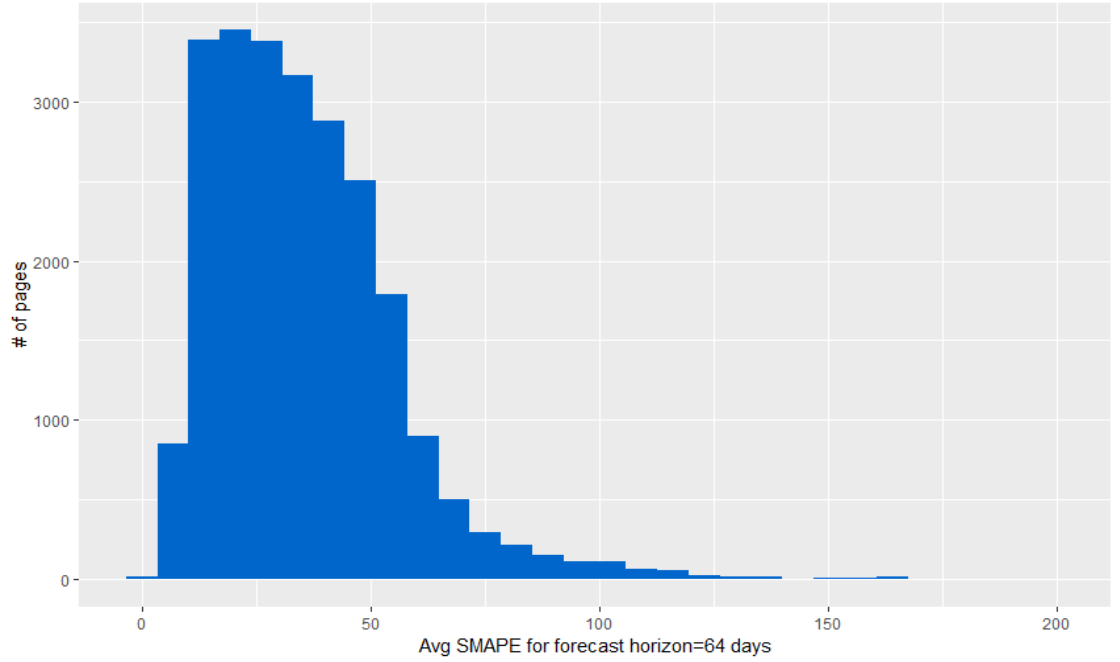
Forecast Horizon: Last 64 days prediction Training Period: 2015-07-01 to 2017-07-08



Avg. SMAPE: 40.55%

Seq. SSA + Conv. Net Approach

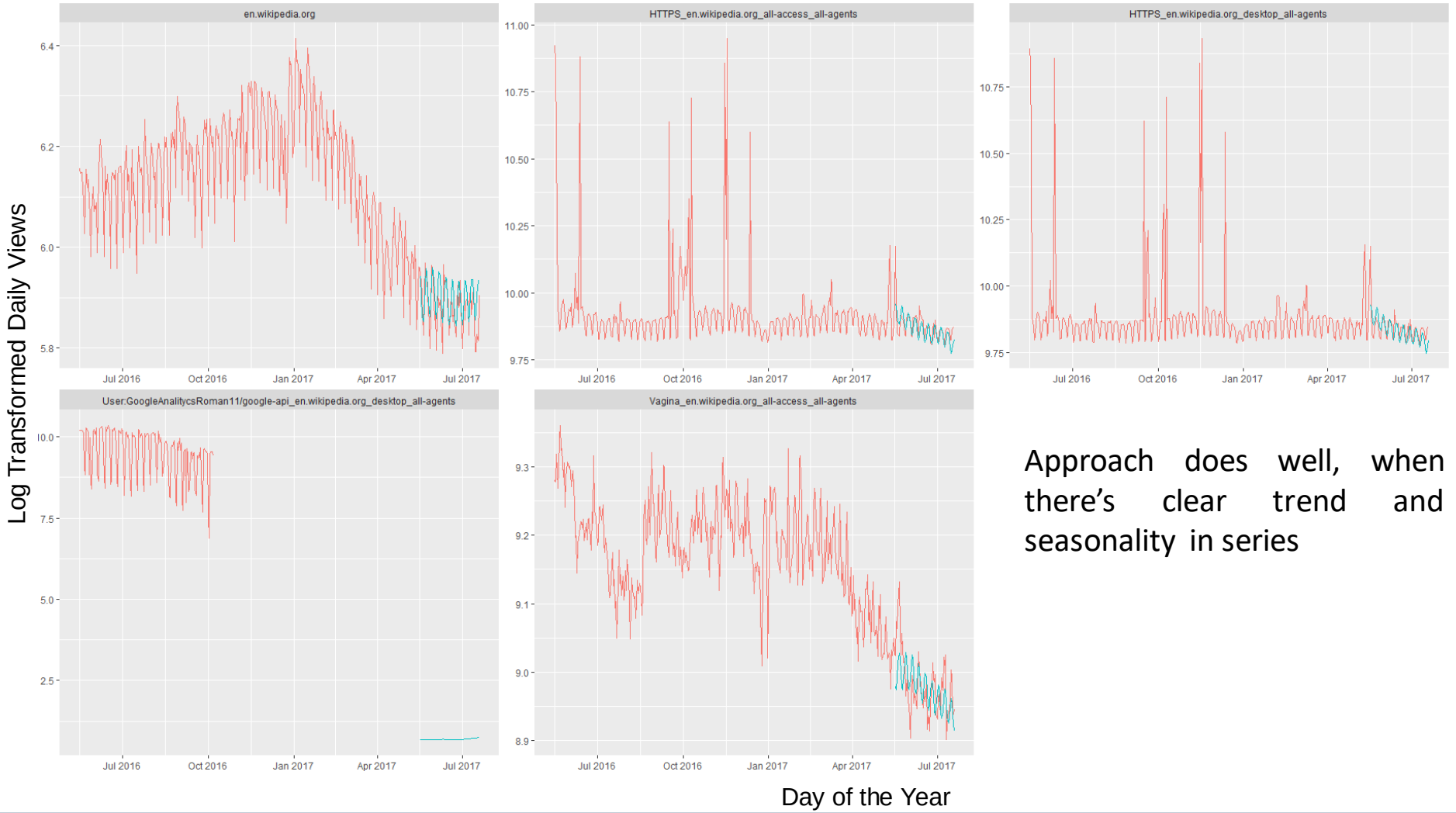
Forecast Horizon: Last 64 days prediction Training Period: 2015-07-01 to 2017-07-08



Avg. SMAPE: 35.68%

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

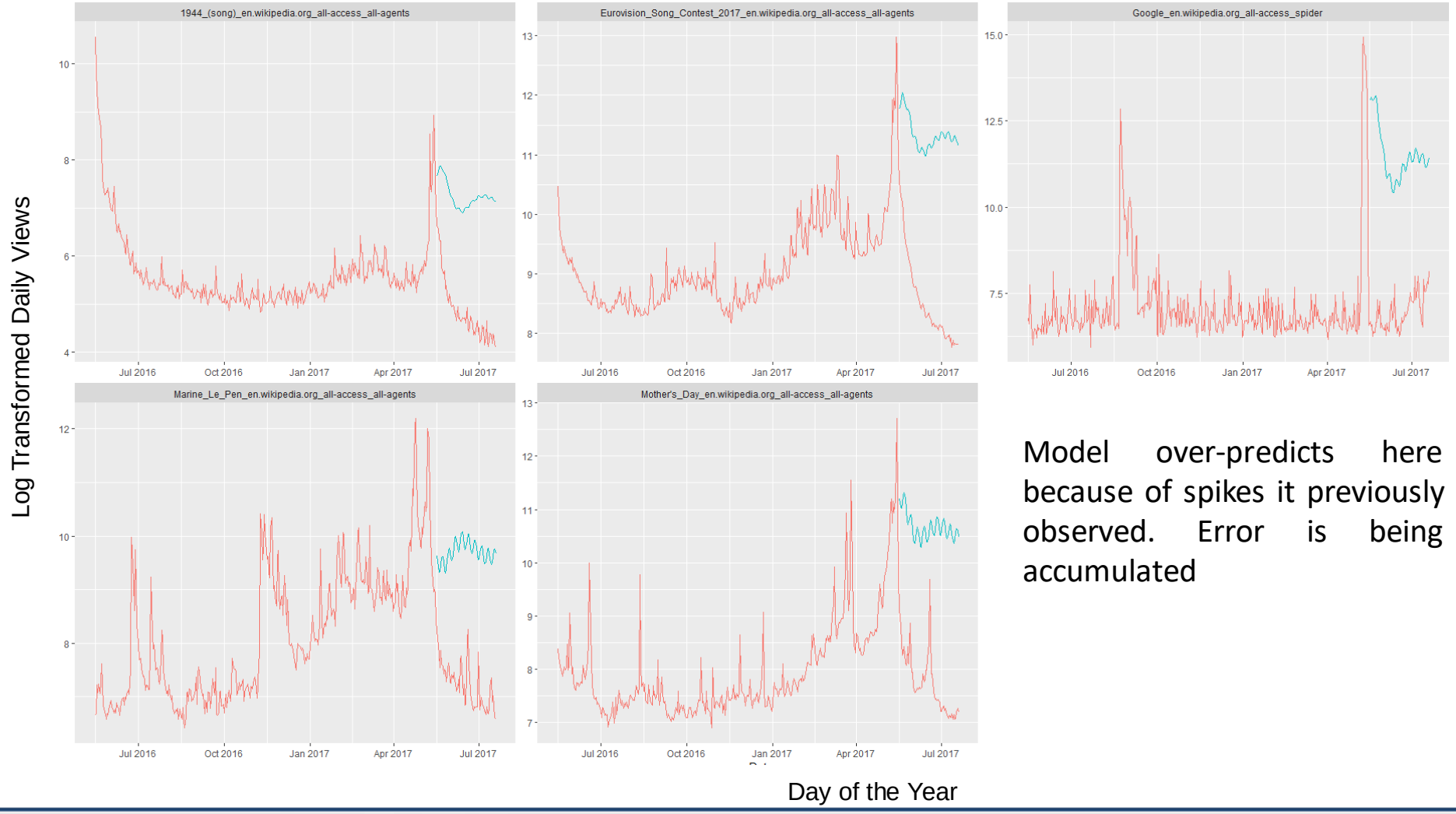
Results and Evaluation- The Good



Approach does well, when there's clear trend and seasonality in series

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

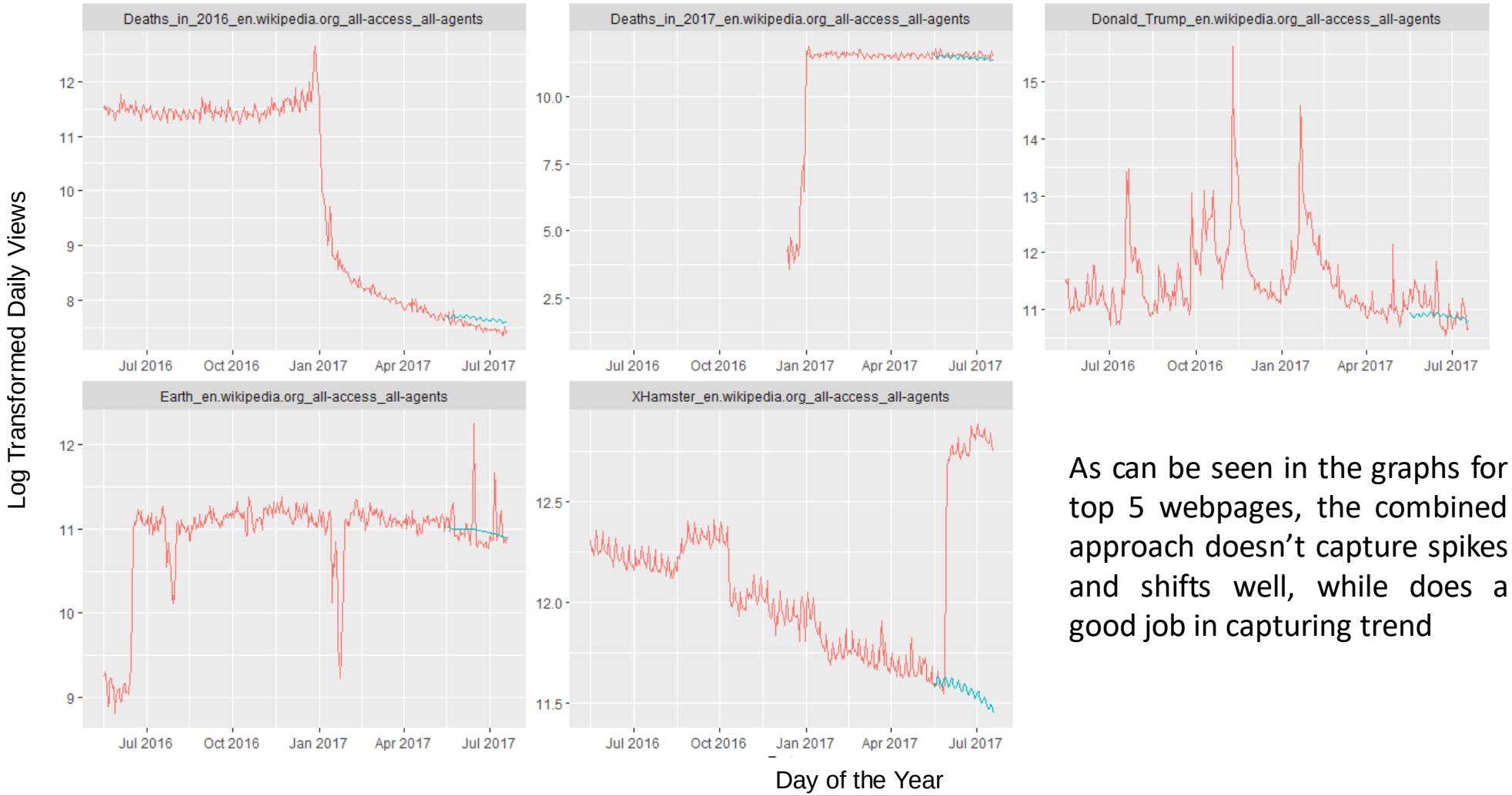
Results and Evaluation- The Bad



Model over-predicts here because of spikes it previously observed. Error is being accumulated

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Results and Evaluation- The Top 5



As can be seen in the graphs for top 5 webpages, the combined approach doesn't capture spikes and shifts well, while does a good job in capturing trend

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Agenda

- ❑Defining the Problem
- ❑Data Description
- ❑Feature Engineering
- ❑Introduction of Methods
- ❑Modelling Approach
- ❑Model Identification
- ❑Results and Evaluation
- ❑Challenges
- ❑Conclusion
- ❑Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Challenges

- Large number of traffic sites in original data ~ 145k series
  - Only a subset of one language project out of 7 was chosen
- Weak correlations among traffic series
  - Univariate approach would have yielded similar results
- Missing data, about 6% of series information is missing, uncertainty in missingness
- Non- stationary series, about 30% of the traffic series is non-stationary
- Scales of traffic differ
- Traditional multivariate and state space models, explode in parameter space with high dimensionality

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Agenda

- ❑Defining the Problem
- ❑Data Description
- ❑Feature Engineering
- ❑Introduction of Methods
- ❑Modelling Approach
- ❑Model Identification
- ❑Results and Evaluation
- ❑Challenges
- ❑Conclusion
- ❑Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Conclusion

- SSA technique is well-suited to decompose a system of non-stationary time-series into matched components
- The window length , L is a parameter that can be tuned
- Sequential SSA, can help identify seasonal effects after detrending the series
- Wave Nets (a CNN network with dilation) can help capture long-range correlations in series by increasing the receptive field of layers using dilation (skipping certain parts of input)
- A combined approach could yield a boosted performance, while maintaining expressiveness

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Agenda

- ❑Defining the Problem
- ❑Data Description
- ❑Feature Engineering
- ❑Introduction of Methods
- ❑Modelling Approach
- ❑Model Identification
- ❑Results and Evaluation
- ❑Data and Modelling Challenges
- ❑Conclusion
- ❑Future Work

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

Future Work

- Oblique and Iterative SSA techniques could be explored to make the decomposition well-separable
- Automate parameter choosing for SSA- window length and grouping
- Tune Wave-Net parameters:
  - Larger kernel size- to see if there's any long range correlations
  - 1X1 Kernel- series is uncorrelated
  - Number of filters and layers
  - Batch size
  - L2 Regularization/ Dropout rate
  - Early Stopping criterion
  - Explore skipping connections, to remove conditioning on uncorrelated series
- Deep Kalman Filters: A time -varying generative model

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu

References

- Dezhi Hong, Quanquan Gu, Kamin Whitehouse. High-dimensional Time Series Clustering via Cross-Predictability. *http://www.cs.virginia.edu/~dh5gm/pdf/aistats17-paper.pdf*
- Nina Golyandina, Anton Korobeynikov, Alex Shlemov, Konstantin Usevich.Multivariate and 2D Extensions of Singular Spectrum Analysis with the Rssa Package. *https://arxiv.org/pdf/1309.5050.pdf*
- Anastasia Borovykh, Sander Bohte, Cornelis W. Oosterlee. Conditional time series forecasting with convolutional neural networks. *https://arxiv.org/pdf/1703.04691.pdf*
- Sokchhay Heng, Tadashi Suetsugi. Coupling Singular Spectrum Analysis with Artificial Neural Network to Improve Accuracy of Sediment Load Prediction. *https://file.scirp.org/pdf/JWARP_20130422145534 28.pdf*
- Nina Golyandina, Anton Korobeynikov, Anatoly Zhigljavsky. A textbook on Singular Spectrum analysis with R.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. WaveNet: A Generative Model for Audio. *https://arxiv.org/pdf/1609.03499.pdf*

Thank You

Bharadwaj Janarthanan | bharadwaj@cs.toronto.edu