Dynamic Poisson Autoregression for Influenza-Like-Illness Case Count Prediction

Zheng Wang[1], Prithwish Chakraborty[2], Sumiko R. Mekaru[3], John S. Brownstein[3], Jieping Ye[1] and Naren Ramakrishnan[2]

[1]University of Michigan, Ann Arbor
[2]Virginia Tech
[3]Boston Children's Hospital, Harvard Medical School

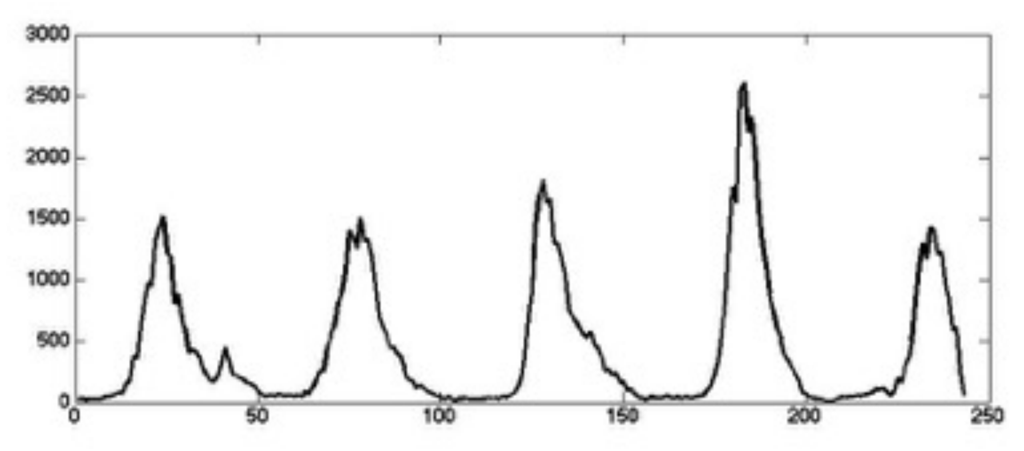
August 13, 2015 @ KDD 2015, Sydney, Australia

Outline

- Introduction
  - Influenza-Like-Illness (ILI) case count prediction
  - Key contributions
- Methodology: formulation and solution
  - Dynamic ARX model
  - Dynamic Poisson ARX model
- Experiments
- Conclusion

Influenza-Like-Illness (ILI) Case Count

- Seasonal influenza regularly affects the global population
- Epidemic diseases forecasting and surveillance
- Case count (#ILI): doctor visit
- Calibrated #ILI over #week

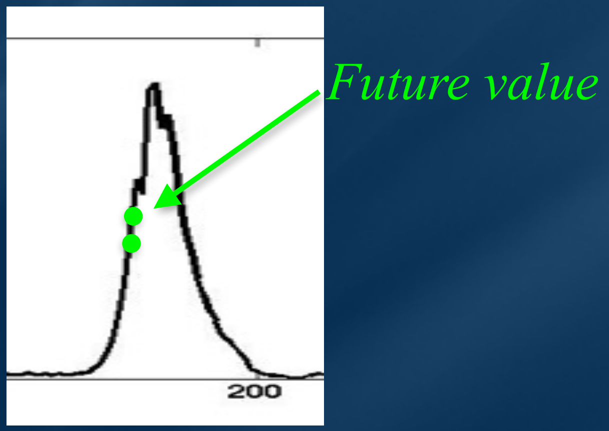


Argentina

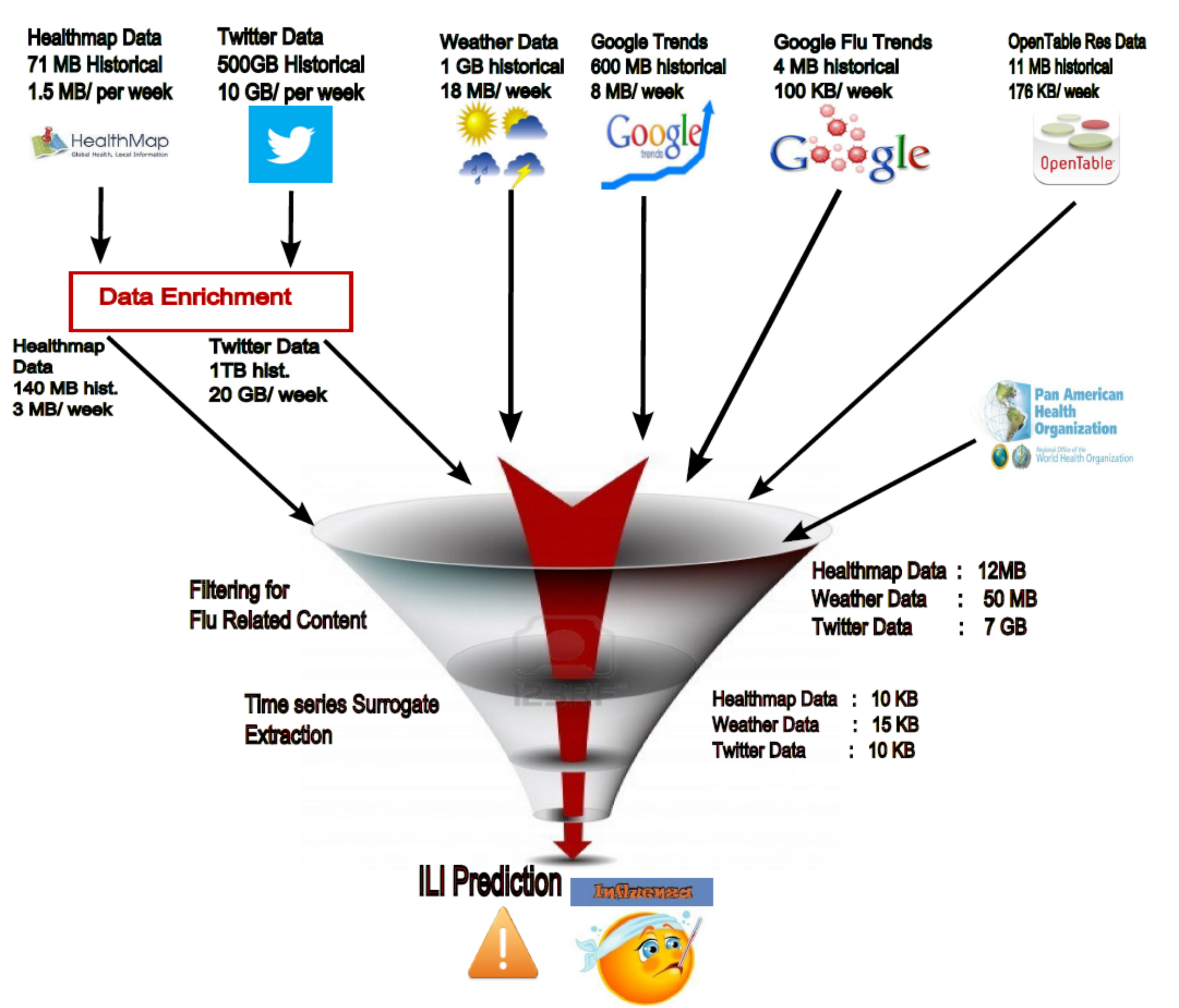
ILI Case Count Prediction

- Long term prediction
  - Season-wise prediction
  - Target: starting time, ending time, peak value, peak time

- Short term prediction
  - Point-wise prediction
  - Target: values for next few time points



ILI Forecast with Indicator Data



The figure is cited from P. Chakraborty et al. *Forecasting a moving target: Ensemble models for ILI case count predictions.* SDM '14, 2014.

Indicator Data Source

- Weather: temperature, humidity
- Social media text data: Twitter, Google search trends(GST)
- Domain data: Paho/CDC (history record), Google flu trends (GFT), HealthMap
- Others: OpenTables …
- Goal: multi-step ILI case count forecasting

Key Contributions

- New dynamic time series prediction model
- Dynamic Poisson ARX model for count data
- Efficient solution with block coordinate descent
- Applicable to other time series forecasting problems

ARX Model

- Autoregressive model with exogenous input

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \boxed{\sum_{i=0}^{b} \beta_i^T \mathbf{x}_{t-d-i}} + \varepsilon_t + c$$

order *p* with input lag *d*

indicator data as the (mutli-dimensional) exogenous input

ARX Model

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{i=0}^{b} \beta_i^T \mathbf{x}_{t-d-i} + \varepsilon_t + c$$

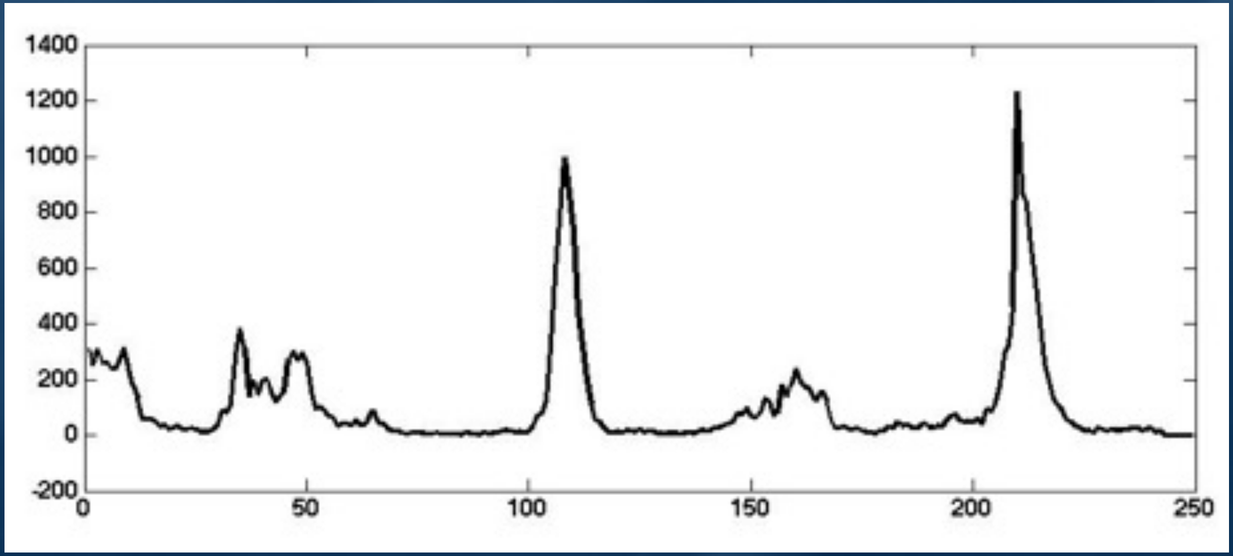
$$\mathbf{w} = \left[\beta_{t-d}, \cdots, \beta_{t-d-b}, \alpha_{t-1}, \cdots, \alpha_{t-p}, c \right] \qquad \mathbf{z}_t = \left[\mathbf{x}_{t-d}, \cdots, \mathbf{x}_{t-d-b}, y_{t-1}, \cdots, y_{t-p}, 1 \right]$$

$$y_t = \mathbf{w}\mathbf{z}_t^T + \varepsilon_t$$

Least Squares Problem: $\min_{\mathbf{w}} \sum_t l(\mathbf{z}_t, y_t) = \sum_t (y_t - \mathbf{w}\mathbf{z}_t^T)^2$

Limitation

- Irregular seasonal pattern in real world



Dynamic Modeling

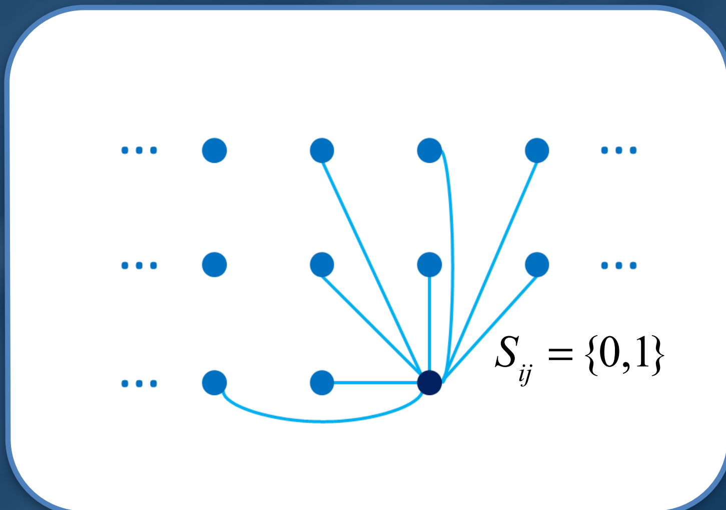
- Time dependent weight: different model for different time point

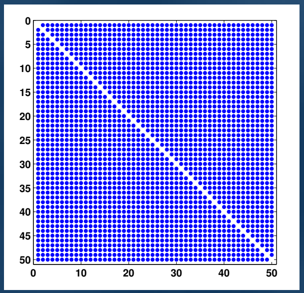
$$y_t = \mathbf{w}_t \mathbf{z}_t^T + \varepsilon_t$$

- Model complexity constraint

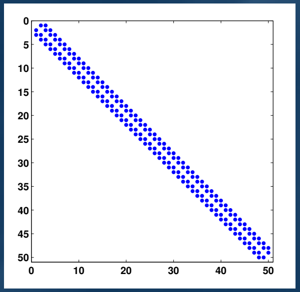
$$\sum_{i=1}^{N} \left(y_i - \mathbf{w}_i \mathbf{z}_i^T \right)^2 + \eta R(\mathbf{w})$$

Similarity Graph

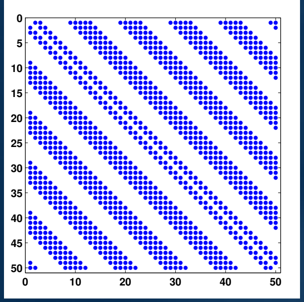




fully connected graph



nearest neighbor graph



seasonal nearest neighbor graph

$$\sum_{i=1}^{N} \left(y_i - \mathbf{w}_i \mathbf{z}_i^T \right)^2 + \eta \sum_{i,j} S_{ij} \left\| \mathbf{w}_i - \mathbf{w}_j \right\|_2^2$$

Solution

- Objective

$$\sum_{i=1}^{N}\left(y_i - \mathbf{w}_i \mathbf{z}_i^T\right)^2 + \eta \sum_{i,j} S_{ij} \left\| \mathbf{w}_i - \mathbf{w}_j \right\|_2^2$$

- Block Coordinate Descent: solve each model by fixing all others alternatively

$$\min_{\mathbf{w}_i}\left(y_i - \mathbf{w}_i \mathbf{z}_i^T\right)^2 + \eta \sum_{j \in B_i} \left\| \mathbf{w}_i - \mathbf{w}_j \right\|_2^2$$

closed-form solution: $\mathbf{w}_i = \left(\mathbf{z}_i^T \mathbf{z}_i + \eta K_i \mathbf{I}\right)^{-1} \left(y_i \mathbf{z}_i^T + \eta \sum_{j \in B_i} \mathbf{w}_j \right)$

Algorithm

Algorithm 1 Dynamic Autoregressive Model with Exogenous Variables (DARX)

input data source $\mathbf{X}$, historical target $\mathbf{y}$.

1: Build the samples $\mathbf{Z}$, initial weight $\mathbf{W}^{(0)}$
2: **repeat**
3: **for** $i = 1, \cdots, N$ **do**
4: Solve sub-problem (5) by

$$\left(\mathbf{z}_i^T \mathbf{z}_i + \eta K_i \mathbf{I}\right)^{-1} \left(y_i \mathbf{z}_i^T + \eta \sum_{j \in B_i} \mathbf{w}_j\right)$$

5: **end for**
6: **until** Terminated

output Regression weight $\mathbf{W}$.

Dynamic Poisson ARX Model

Poisson distribution

$$\Pr\left(y\right)=\frac{\lambda^{y}e^{-\lambda}}{y!}$$

Link function

$$\log\left(\mathbb{E}[y\mid\mathbf{z}]\right)=\log(\lambda)=\mathbf{w}\mathbf{z}^{T}$$

Optimization Problem (maximizing log-likelihood):

$$\begin{aligned}\min_{\mathbf{w}}\quad & \sum_{i}\left(\mathbf{w}_i\mathbf{z}_i^{T}-y_i\log(\mathbf{w}_i\mathbf{z}_i^{T})\right)+\eta\sum_{i,j}S_{ij}\left\|\mathbf{w}_i-\mathbf{w}_j\right\|_2^2\\ s.t.\quad & \mathbf{w}_i\mathbf{z}_i^{T}\geq 0,\ \forall i.\end{aligned}$$

Solution

Block Coordinate Descent

$$\min_{\mathbf{w}_i} \quad \left(\mathbf{w}_i\mathbf{z}_i^T - y_i \log(\mathbf{w}_i\mathbf{z}_i^T)\right) + \eta \sum_j S_{ij} \left\|\mathbf{w}_i - \mathbf{w}_j\right\|_2^2$$

$$s.t. \quad \mathbf{w}_i\mathbf{z}_i^T \geq 0.$$

Each subproblem is solved by Newton-Raphson method

$$\mathbf{w}_i \leftarrow \mathbf{w}_i - \mathbf{H}_i^{-1}\mathbf{g}_i \qquad \mathbf{g}_i = \left(1 - \frac{y_i}{\mathbf{w}_i\mathbf{z}_i^T}\right)\mathbf{z}_i + 2\eta \sum_j S_{ij}\left(\mathbf{w}_i - \mathbf{w}_j\right)$$

$$\mathbf{w}_i \leftarrow \mathbf{w}_i - \frac{\mathbf{w}_i\mathbf{z}_i^T}{\mathbf{z}_i\mathbf{z}_i^T}\mathbf{z}_i \qquad \mathbf{H}_i = \frac{y_i}{\left(\mathbf{w}\mathbf{z}_i^T\right)^2}\left(\mathbf{z}_i^T\mathbf{z}_i\right) + 2\eta \sum_j S_{ij}$$

Experiments

Table 1: Data source characteristics. Delayed refers to whether the data source is available for a given week in the same week or later. Revised refers to whether older values can get revised in future updates.

Characteristics	Num. Dimensions	Delayed?	Revised?	Temporal Resolution	Spatial Resolution
PAHO/CDC	1	Yes	Yes	Weekly	Country
Google Flue Trends (GFT)	1	No	Yes	Weekly	Country
Google Search Trends (GST)	114	No	Yes	Weekly	Country
Weather	5	No	No	6 hours $\rightarrow$ Weekly	4 locations $\rightarrow$ Country
HealthMap	114×3	No	No	Daily $\rightarrow$ Weekly	Country

Datasets: United States (US) and 14 Latin American (LA) countries including Argentina (AR), Bolivia (BO), Chile (CL), Colombia (CO), Costa Rica (CR), Ecuador (EC), Guatemala (GT), Honduras (HN), Mexico (MX), Nicaragua (NI), Panama (PA), Peru (PE), Paraguay (PY) and, El Salvador (SV)

Algorithms: ARX, NMF, SARX, **DARX**, **DPARX**

Measures:

$$error = \frac{4}{N}\sum_{t=t_s}^{t_e}\frac{|y_t - \hat{y}_t|}{\max(y_t, \hat{y}_t, 10)} \qquad acc = 4 - error \in [0,4]$$

Prediction Accuracy

Table 2. Prediction accuracies for competing algorithms with different forecast steps over different countries using the GFT input source.

Step	**Method**	AR	BO	CL	MX	PE	PY	US
1	ARX	2.85	2.63	3.18	2.61	2.51	2.82	3.71
	MFN	2.33	2.41	2.34	2.69	2.48	2.54	3.73
	SARX	3.02	2.42	3.11	2.90	2.81	2.69	3.67
	DARX	3.05	2.74	3.12	2.78	2.50	2.65	3.71
	DPARX	3.13	2.82	3.18	2.97	2.64	2.81	3.72
2	ARX	2.38	2.22	2.83	1.88	1.90	2.57	3.47
	MFN	2.12	2.00	2.13	2.33	2.21	2.19	3.63
	SARX	2.75	2.03	2.76	2.64	2.43	2.43	3.64
	DARX	2.94	2.68	3.02	2.58	2.38	2.58	3.60
	DPARX	2.86	2.70	2.89	2.64	2.52	2.65	3.61
3	ARX	2.11	1.86	2.61	1.28	1.44	2.31	3.19
	MFN	1.99	1.87	2.11	2.14	2.10	2.09	3.33
	SARX	2.33	1.61	2.46	2.42	2.16	2.23	3.40
	DARX	2.66	2.36	2.77	2.37	2.26	2.46	3.41
	DPARX	2.58	2.53	2.56	2.45	2.37	2.52	3.42
4	ARX	1.84	1.61	2.39	0.88	1.12	2.22	2.92
	MFN	1.85	1.83	2.00	2.05	2.01	1.94	3.15
	SARX	2.12	1.41	2.30	2.22	2.02	2.09	3.30
	DARX	2.34	2.21	2.52	1.98	2.19	2.22	3.18
	DPARX	2.29	2.35	2.32	2.26	2.29	2.40	3.20

Prediction accuracy on other input sources

Step	Method	AR	BO	CL	CO	CR	EC	GT	HN	MX	NI	PA	PE	PY	SV	US
1	ARX	2.94	2.51	3.10	2.90	2.21	2.81	2.83	2.96	2.25	2.18	2.78	2.51	2.84	2.83	3.51
	MFN	2.99	3.01	2.88	2.53	2.78	2.81	2.77	2.83	2.61	2.70	2.56	2.82	2.66	2.79	3.81
	DARX	3.09	2.84	3.17	2.84	2.57	2.94	2.83	2.89	2.91	2.77	2.72	2.67	2.79	2.72	3.71
	DPARX	2.98	2.84	3.07	3.01	2.70	2.97	2.87	2.93	2.84	2.86	2.82	2.78	2.86	2.77	3.72
2	ARX	2.56	2.05	2.63	2.71	1.61	2.56	2.63	2.76	1.15	1.36	2.56	2.05	2.62	2.64	3.21
	MFN	2.86	2.89	2.81	2.49	2.71	2.67	2.72	2.41	2.55	2.31	2.50	2.59	2.71	2.30	3.75
	DARX	2.98	2.69	3.00	2.69	2.63	2.79	2.72	2.81	2.66	2.28	2.55	2.49	2.68	2.66	3.60
	DPARX	2.67	2.73	2.86	2.83	2.66	2.79	2.78	2.78	2.62	2.49	2.71	2.63	2.64	2.68	3.61
3	ARX	2.25	1.65	2.21	2.50	1.06	2.30	2.39	2.59	0.60	0.94	2.42	1.72	2.39	2.46	2.92
	MFN	2.49	2.38	2.41	2.33	2.45	2.31	2.32	2.10	2.21	2.11	2.19	2.22	2.40	2.08	3.64
	DARX	2.68	2.32	2.68	2.57	2.52	2.72	2.50	2.65	2.47	2.00	2.52	2.32	2.54	2.53	3.41
	DPARX	2.33	2.44	2.63	2.70	2.58	2.66	2.59	2.61	2.36	2.31	2.75	2.44	2.51	2.55	3.42
4	ARX	1.98	1.37	1.73	2.31	0.72	2.07	2.22	2.41	0.39	0.83	2.21	1.46	2.21	2.30	2.56
	MFN	2.10	2.13	2.15	2.04	2.25	2.11	2.22	1.94	1.99	1.87	2.01	1.86	2.10	1.77	3.54
	DARX	2.42	2.12	2.39	2.49	2.34	2.52	2.42	2.51	2.17	1.74	2.38	2.27	2.30	2.42	3.18
	DPARX	2.10	2.23	2.32	2.64	2.38	2.52	2.55	2.45	2.06	2.15	2.72	2.38	2.27	2.53	3.20

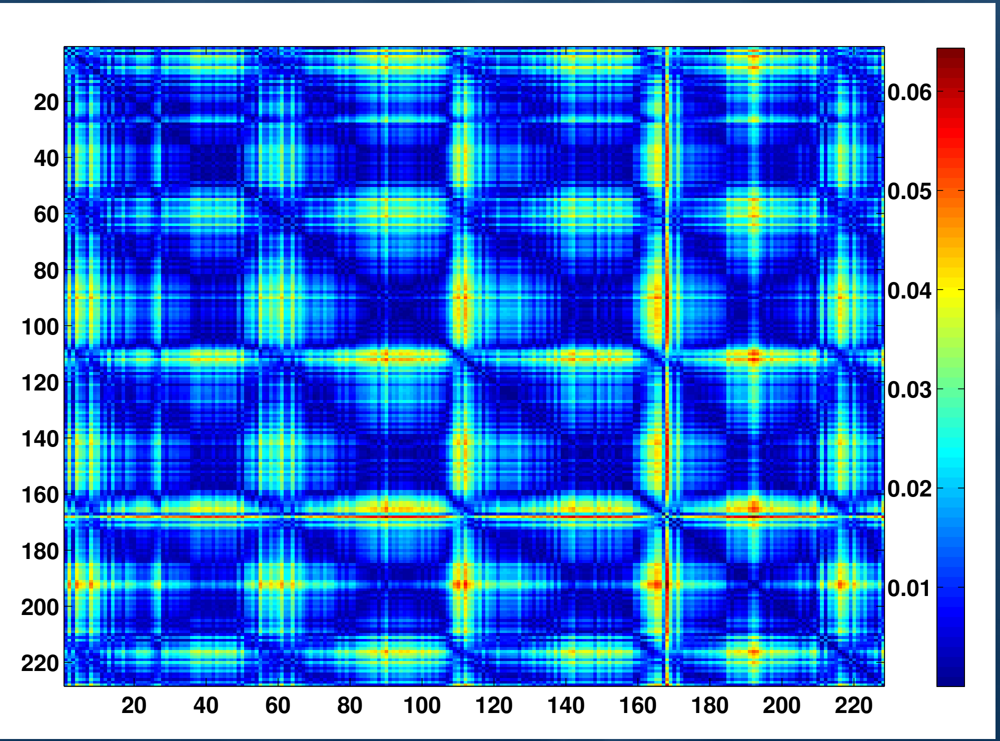
Step	Dataset	AR	BO	CL	CO	CR	EC	GT	HN	MX	NI	PA	PE	PY	SV
1	MFN	2.61	2.44	2.55	2.22	2.61	2.52	2.31	2.62	2.48	2.61	2.31	2.23	2.53	2.13
	DARX	2.99	2.65	3.09	2.74	2.41	2.86	2.72	2.83	2.82	2.84	2.59	2.56	2.75	2.63
	DPARX	3.07	2.74	3.15	2.85	2.72	2.80	2.51	2.80	2.96	2.77	2.59	2.66	2.82	2.61
2	MFN	2.50	2.33	2.31	2.10	2.44	2.29	2.11	2.43	2.37	2.39	2.20	2.01	2.27	2.00
	DARX	2.83	2.54	2.94	2.57	2.53	2.69	2.58	2.72	2.59	2.40	2.35	2.40	2.54	2.51
	DPARX	2.78	2.59	2.86	2.67	2.63	2.67	2.35	2.71	2.60	2.48	2.43	2.53	2.57	2.59
3	MFN	2.33	2.10	2.16	1.99	2.21	2.03	1.99	2.14	2.20	2.14	2.02	1.91	2.13	1.92
	DARX	2.51	2.07	2.69	2.45	2.36	2.47	2.41	2.54	2.34	2.06	2.48	2.10	2.49	2.44
	DPARX	2.46	2.41	2.53	2.56	2.48	2.51	2.26	2.58	2.38	2.30	2.41	2.34	2.49	2.51
4	MFN	1.99	2.00	2.01	1.82	1.97	1.88	1.92	1.93	1.81	1.77	1.79	1.70	1.82	1.71
	DARX	2.16	1.91	2.36	2.24	2.20	2.17	2.28	2.40	1.80	1.86	2.40	2.06	2.23	2.36
	DPARX	2.17	2.21	2.29	2.46	2.35	2.33	2.14	2.46	2.10	2.13	2.33	2.21	2.30	2.44

Step	Dataset	AR	BO	CL	CO	CR	EC	GT	HN	MX	NI	PA	PE	PY	SV	US
1	MFN	2.81	3.13	2.63	2.58	2.91	2.77	2.63	2.73	2.50	2.61	2.54	2.69	2.51	2.61	3.78
	DARX	3.00	2.69	3.11	2.79	2.44	2.89	2.75	2.91	2.85	2.86	2.60	2.65	2.75	2.64	3.71
	DPARX	3.07	2.74	3.15	2.84	2.69	2.83	2.58	2.82	2.95	2.79	2.59	2.70	2.83	2.62	3.72
2	MFN	2.71	2.91	2.30	2.21	2.77	2.49	2.40	2.38	2.44	2.36	2.15	2.33	2.22	2.33	3.64
	DARX	2.86	2.60	3.01	2.62	2.54	2.74	2.64	2.77	2.66	2.47	2.37	2.47	2.53	2.58	3.60
	DPARX	2.78	2.60	2.88	2.67	2.62	2.71	2.44	2.72	2.60	2.50	2.45	2.58	2.58	2.60	3.61
3	MFN	2.44	2.30	2.42	2.07	2.31	2.14	2.28	2.01	2.19	2.12	1.99	2.00	1.97	1.95	3.35
	DARX	2.58	2.18	2.78	2.49	2.35	2.63	2.51	2.62	2.48	2.15	2.49	2.33	2.48	2.51	3.41
	DPARX	2.46	2.42	2.55	2.56	2.47	2.58	2.36	2.59	2.38	2.31	2.45	2.37	2.49	2.50	3.42
4	MFN	1.93	1.99	2.20	1.88	2.00	1.95	2.15	1.95	1.89	1.85	1.72	1.78	1.91	1.81	3.13
	DARX	2.28	2.02	2.46	2.39	2.19	2.37	2.39	2.45	2.22	1.97	2.45	2.26	2.20	2.42	3.18
	DPARX	2.17	2.21	2.30	2.44	2.34	2.42	2.25	2.47	2.12	2.14	2.37	2.25	2.30	2.47	3.21

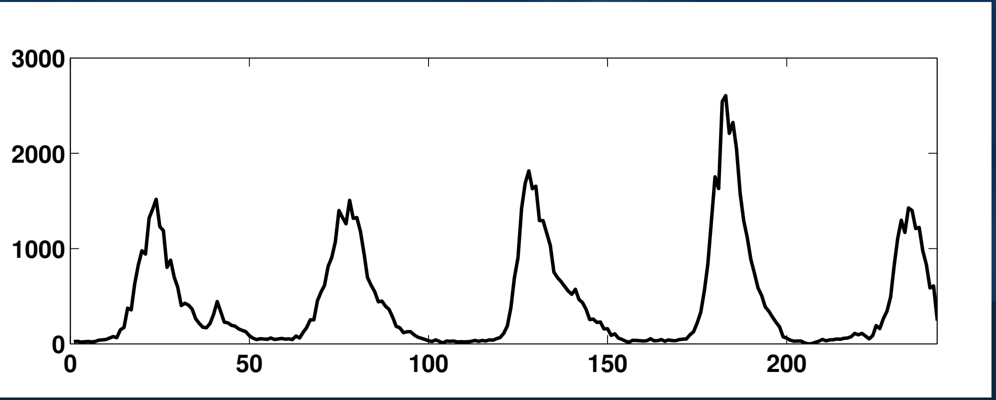
Model Similarity

model distance matrix

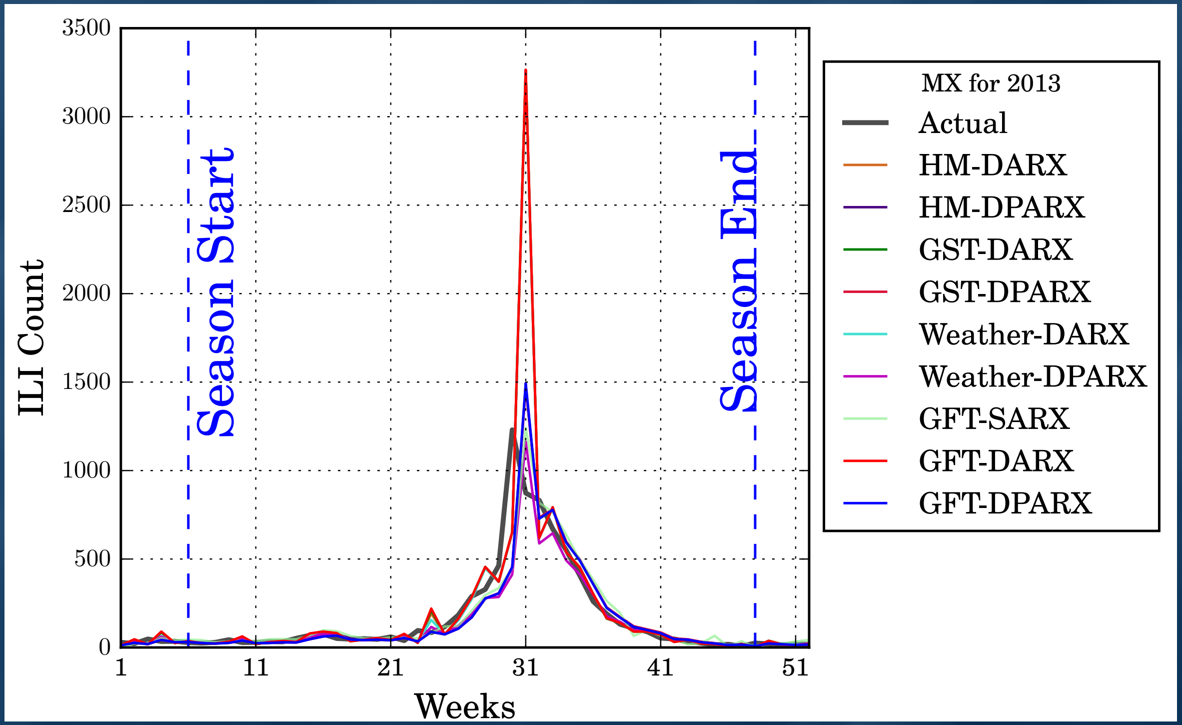
$$d_{ij} = \left\| \mathbf{w}_i - \mathbf{w}_j \right\|_2$$



ILI case count series



Long Term Prediction



Comparison of seasonal characteristics for Mexico using different algorithms for one-step ahead prediction. Blue vertical dashed lines indicate the actual start and end of the season. ILI season considered: 2013.

Conclusion

- ILI case count prediction is an important time series prediction problem.
- Limitation of the conventional time series model.
- Dynamic model is more appropriate.
- Further work: fuse different indicator data sources together.

Thanks!