Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions

Prithwish Chakraborty[1,2], Pejman Khadivi[1,2], Bryan Lewis[3], Aravindan Mahendiran[1,2], Jiangzhuo Chen[3], Patrick Butler[1,2], Elaine O. Nsoesie[3,4,5], Sumiko R. Mekaru[4,5], John S. Brownstein[4,5], Madhav V. Marathe[3], Naren Ramakrishnan[1,2]

[1]Dept. of Computer Science, Virginia Tech, USA
[2]Discovery Analytics Center, Virginia Tech, USA
[3]NDSSL, Virginia Bioinformatics Institute, USA
[4]Children's Hopsital Informatics Program, Boston Children's Hopsital, USA
[5]Dept. of Pediatrics, Harvard Medical School, USA.

October 26, 2014

Problem Overview

Problem Overview

- Predicting weekly Influenza-like-illness (ILI) case counts for 15 Latin American countries
- Investigating different open source data-streams as possible surrogate indicators of ILI

Motivation

- Traditional methods are often not enough!!
  - ILI surveillance is not real-time - often lags several weeks
  - Estimates are “unstable” - often revised over several months

Motivation

- Traditional methods are often not enough!!
  - ILI surveillance is not real-time - often lags several weeks
  - Estimates are "unstable" - often revised over several months
- Can surrogate information be used to provide more stable and real time estimates?
  - Either "non-physical indicators" or "physical indicators" investigated
  - How to handle the instability associated with ILI surveillance

.

Key Contributions

1. Real-time prospective study - most studies till now have been retrospective

Key Contributions

1. Real-time prospective study - most studies till now have been retrospective
2. Integrates both social and physical indicators

Key Contributions

1. Real-time prospective study - most studies till now have been retrospective
2. Integrates both social and physical indicators
3. Data level fusion vs Model level fusion?

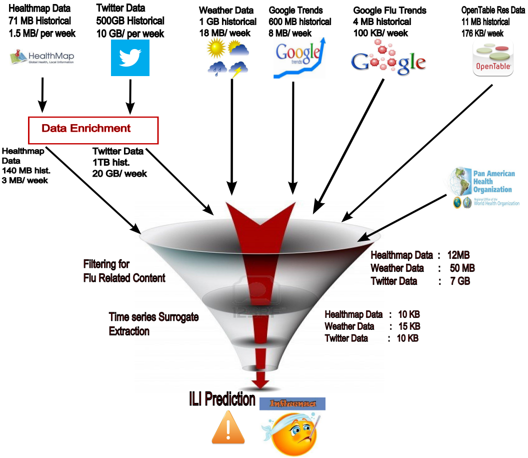
Key Contributions

1. Real-time prospective study - most studies till now have been retrospective
2. Integrates both social and physical indicators
3. Data level fusion vs Model level fusion?
4. Accounting for uncertainties in the official surveillance estimates

Key Contributions

1. Real-time prospective study - most studies till now have been retrospective
2. Integrates both social and physical indicators
3. Data level fusion vs Model level fusion?
4. Accounting for uncertainties in the official surveillance estimates
5. Investigate importance of different sources - Ablation test

Overall Framework



Data Sources

- Non-physical indicators

Data Sources

- Non-physical indicators
  1. Google Flu Trends - uses unpublished set of keywords

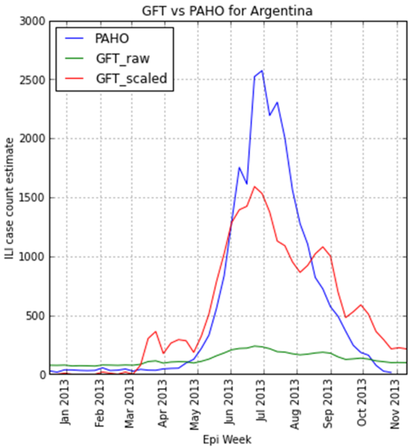
Data Sources

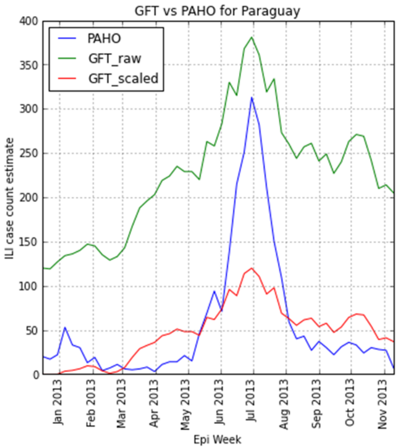
- Non-physical indicators
  1. Google Flu Trends - uses unpublished set of keywords
  2. Custom User Keywords
     1. Google Search Trends
     2. Healthmap News Feed
     3. Twitter Feed

Data Sources

- Non-physical indicators
  1. Google Flu Trends - uses unpublished set of keywords
  2. Custom User Keywords
     1. Google Search Trends
     2. Healthmap News Feed
     3. Twitter Feed
- Physical indicators
- Misc. Indicators
  1. Opentable reservations

Google Flu Trends





Finding Custom user keyword dictionary

A multiple step process :

- Started with a seed set of keywords from experts.
  - Seed set contains words in Spanish, Portuguese, and English.
  - example : *gripe* (flu in Spanish)

Finding Custom user keyword dictionary

A multiple step process :

- Started with a seed set of keywords from experts.
  - Seed set contains words in Spanish, Portuguese, and English.
  - example : *gripe* (flu in Spanish)
- Pseudo-query expansion
  - Crawled top 20 web-sites for each seed word.
  - Crawled "expert" web-sites e.g. CDC.
  - Crawled few other hand-picked sites.
  - Top 500 frequently occurring words selected.

Finding Custom user keyword dictionary

A multiple step process :

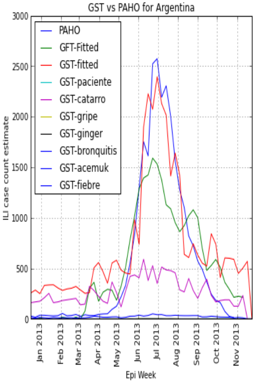
- Started with a seed set of keywords from experts.
  - Seed set contains words in Spanish, Portuguese, and English.
  - example : *gripe* (flu in Spanish)
- Pseudo-query expansion
  - Crawled top 20 web-sites for each seed word.
  - Crawled "expert" web-sites e.g. CDC.
  - Crawled few other hand-picked sites.
  - Top 500 frequently occurring words selected.
- Time series correlation analysis
  - Used Google Correlate to find words with search history correlated with ILI incidence curve.
  - Interesting words such as *ginger* and *Acemuk* found.

Finding Custom user keyword dictionary

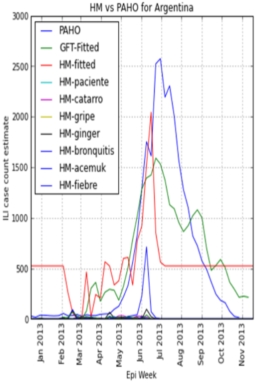
A multiple step process :

- Started with a seed set of keywords from experts.
  - Seed set contains words in Spanish, Portuguese, and English.
  - example : *gripe* (flu in Spanish)
- Pseudo-query expansion
  - Crawled top 20 web-sites for each seed word.
  - Crawled "expert" web-sites e.g. CDC.
  - Crawled few other hand-picked sites.
  - Top 500 frequently occurring words selected.
- Time series correlation analysis
  - Used Google Correlate to find words with search history correlated with ILI incidence curve.
  - Interesting words such as *ginger* and *Acemuk* found.
- Final filtering : 114 words

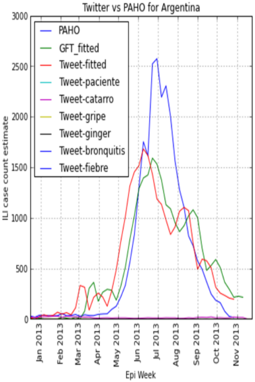
GFT vs other non-physical indicators using custom keyword set



Google Search Trends (GST)



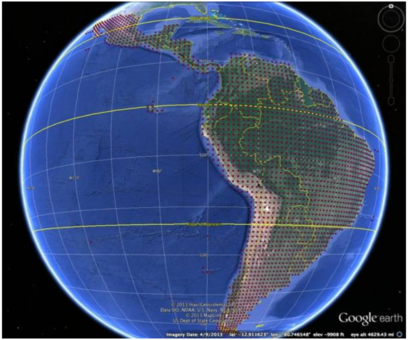
Healthmap



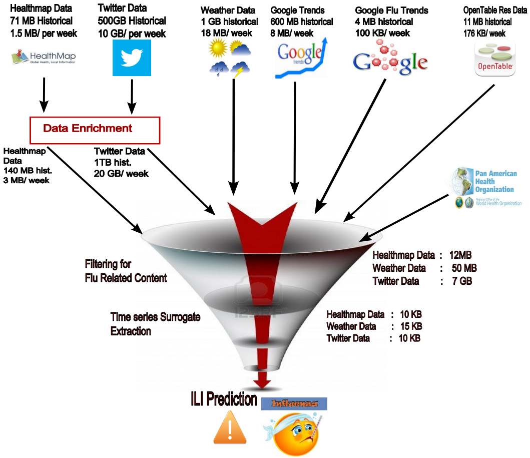
Twitter

Physical Indicators

- Meteorological data for every lat-long, worldwide, every 8 hours
- Humidity, Temperature, Rainfall
- Analyzing grid cells covering PAHO sites.



System framework once again!!



Preliminaries

- To find predictive model f

$$f : \mathcal{P}_t = f\left(\mathcal{P}, \mathcal{X}\right)$$

- Variable Setup

$$\begin{aligned} V_t &\equiv \langle P_{t-\beta-\alpha}, \mathcal{X}_{t-\beta-\alpha}, P_{t+1-\beta-\alpha}, \mathcal{X}_{t+1-\beta-\alpha}, \ldots, \\ &\quad P_{t-\alpha}, \mathcal{X}_{t-\alpha} \rangle \\ L_t &\equiv P_t \end{aligned}$$

- Parameters
  - α : the *lookahead window length*
  - β : the *lookback window length*

Matrix Factorization (MF)

- Can find latent factors in the dataset.

Matrix Factorization (MF)

- Can find latent factors in the dataset.
- Model

$$\widehat{\mathcal{M}}_{i,j} = b_{u,i} + U_i^T F_j$$
$$b_{i,j} = \bar{\mathcal{M}} + b_j$$

Matrix Factorization (MF)

- Can find latent factors in the dataset.
- Model

$$\widehat{\mathcal{M}}_{i,j} = b_{u,i} + U_i^T F_j$$
$$b_{i,j} = \bar{\mathcal{M}} + b_j$$

- Fitting

$$\begin{aligned} b_*, F, U = \operatorname{argmin}(& \sum_{i=1}^{m-1} \left(\mathcal{M}_{i,n} - \widehat{\mathcal{M}}_{i,n}\right)^2 \\ & + \lambda_1 (\sum_{j=1}^{n} b_j^2 + \sum_{i=1}^{m-1} ||U_i||^2 + \sum_{j=1}^{n} ||F_j||^2)) \end{aligned} \tag{1}$$

Nearest Neighbor model (NN)

- Impose non-linearity.

Nearest Neighbor model (NN)

- Impose non-linearity.
- $\mathcal{N}(i) = \{k : V_k \text{ is one of the top K nearest neighbors of } V_i\}$

Nearest Neighbor model (NN)

- Impose non-linearity.
- $\mathcal{N}(i) = \{k : V_k \text{ is one of the top K nearest neighbors of } V_i\}$
- Fitting

$$\widehat{P}_{T'} = \Big(\sum_{k \in \mathcal{N}(T')} \theta_k L_{k,T-\alpha}\Big) \Big/ \sum_{k=1}^{K} \theta_k \tag{2}$$

Matrix Factorization using Nearest Neighborhood (MFN)

- Inspired from Koren et al.'s work* in Recommender systems.

Matrix Factorization using Nearest Neighborhood (MFN)

- Inspired from Koren et al.'s work* in Recommender systems.
- $$\begin{aligned}\widehat{\mathcal{M}}_{i,j} &= b_{i,j} + U_i^T F_j \\ &+ F_j |\mathcal{N}(i)|^{-\frac{1}{2}} \textstyle\sum_{k \in N(i)} (\mathcal{M}_{i,k} - b_{i,k}) x_k\end{aligned} \tag{3}$$

Matrix Factorization using Nearest Neighborhood (MFN)

- Inspired from Koren et al.'s work* in Recommender systems.
- $$\begin{aligned} \widehat{\mathcal{M}}_{i,j} &= b_{i,j} + U_i^T F_j \\ &+ F_j |\mathcal{N}(i)|^{-\frac{1}{2}} \textstyle\sum_{k \in N(i)} (\mathcal{M}_{i,k} - b_{i,k}) x_k \end{aligned} \tag{3}$$
- Fitting

$$\begin{aligned} b_*, F, U, x_* &= \operatorname{argmin}(\sum_{i=1}^{m-1} \left(\mathcal{M}_{i,n} - \widehat{\mathcal{M}}_{i,n}\right)^2 \\ &+ \lambda_2 (\sum_{j=1}^{n} b_j^2 + \sum_{i=1}^{m-1} ||U_i||^2 + \sum_{j=1}^{n} ||F_j||^2 + \sum_{k} ||x_k||^2)) \end{aligned} \tag{4}$$

* **koren2008factor**

Accuracy comparison

- Quality Metric

$$\mathcal{A} = \frac{4}{N_p} \sum_{t=t_s}^{t_e} \left(1 - \frac{|P_t - \hat{P}_t|}{max(P_t, \hat{P}_t, 10)} \right) \tag{5}$$

Accuracy comparison

Table 1: Comparing forecasting accuracy of models using individual sources. Scores in this and other tables are normalized to [0,4] so that 4 is the most accurate.

Model	Sources	AR	BO	CL	CR	CO	EC	GF	GT	HN	MX	NI	PA	PY	PE	SV	All
MF	$\mathcal{W}$	2.78	2.46	2.39	2.14	2.70	2.22	2.12	2.63	2.52	**2.73**	2.31	2.21	2.49	2.77	**2.61**	2.47
	$\mathcal{H}$	2.81	2.31	2.22	1.92	2.43	2.04	2.11	2.57	2.33	2.48	2.39	2.15	2.18	2.47	2.33	2.32
	$\mathcal{T}$	2.37	2.35	2.18	2.03	2.21	2.12	1.83	2.12	2.29	2.03	1.89	2.06	1.96	2.20	2.21	2.12
	$\mathcal{F}$	2.34	2.11	2.29	N/A	N/A	N/A	N/A	N/A	N/A	2.71	N/A	N/A	2.31	2.24	N/A	2.33
	$\mathcal{S}$	2.48	2.21	2.33	2.04	2.31	2.21	1.93	2.03	2.15	2.51	2.42	2.52	2.33	1.93	2.30	2.24
NN	$\mathcal{W}$	2.92	2.93	2.63	2.52	2.66	2.51	2.71	2.82	2.59	2.62	2.55	2.59	2.61	**2.80**	2.52	2.66
	$\mathcal{H}$	2.73	3.10	2.42	2.27	2.83	2.64	2.43	2.25	2.71	2.31	**2.61**	2.35	2.43	2.39	2.52	2.53
	$\mathcal{T}$	2.72	2.86	2.31	2.62	2.77	2.52	2.71	2.66	2.51	2.44	2.13	2.01	1.77	2.51	2.20	2.45
	$\mathcal{F}$	2.11	2.21	2.33	N/A	N/A	N/A	N/A	N/A	N/A	2.19	N/A	N/A	2.41	2.32	N/A	2.26
	$\mathcal{S}$	2.51	2.31	2.41	1.81	2.52	2.41	2.12	2.29	2.51	2.13	**2.61**	2.14	2.51	1.87	2.12	2.28
MFN	$\mathcal{W}$	**2.99**	3.01	**2.88**	2.53	2.78	**2.81**	2.77	**2.83**	2.61	2.70	2.56	**2.66**	**2.82**	2.79	2.51	**2.75**
	$\mathcal{H}$	2.81	**3.13**	2.63	2.58	**2.91**	2.77	2.57	2.63	**2.73**	2.50	**2.61**	2.54	2.51	2.69	**2.61**	2.68
	$\mathcal{T}$	2.74	3.03	2.51	**2.64**	2.83	2.51	**2.81**	2.71	2.60	2.48	2.13	2.55	2.19	2.57	2.31	2.57
	$\mathcal{F}$	2.33	2.41	2.34	N/A	N/A	N/A	N/A	N/A	N/A	2.69	N/A	N/A	2.54	2.48	N/A	2.46
	$\mathcal{S}$	2.61	2.44	2.55	2.22	2.61	2.52	2.71	2.31	2.62	2.48	**2.61**	2.31	2.53	2.23	2.13	2.46

Accuracy comparison

Table 1: Comparing forecasting accuracy of models using individual sources. Scores in this and other tables are normalized to [0,4] so that 4 is the most accurate.

Model	Sources	AR	BO	CL	CR	CO	EC	GF	GT	HN	MX	NI	PA	PY	PE	SV	All
MF	$\mathcal{W}$	2.78	2.46	2.39	2.14	2.70	2.22	2.12	2.63	2.52	**2.73**	2.31	2.21	2.49	2.77	**2.61**	2.47
	$\mathcal{H}$	2.81	2.31	2.22	1.92	2.43	2.04	2.11	2.57	2.33	2.48	2.39	2.15	2.18	2.47	2.33	2.32
	$\mathcal{T}$	2.37	2.35	2.18	2.03	2.21	2.12	1.83	2.12	2.29	2.03	1.89	2.06	1.96	2.20	2.21	2.12
	$\mathcal{F}$	2.34	2.11	2.29	N/A	N/A	N/A	N/A	N/A	N/A	2.71	N/A	N/A	2.31	2.24	N/A	2.33
	$\mathcal{S}$	2.48	2.21	2.33	2.04	2.31	2.21	1.93	2.03	2.15	2.51	2.42	2.52	2.33	1.93	2.30	2.24
NN	$\mathcal{W}$	2.92	2.93	2.63	2.52	2.66	2.51	2.71	2.82	2.59	2.62	2.55	2.59	2.61	**2.80**	2.52	2.66
	$\mathcal{H}$	2.73	3.10	2.42	2.27	2.83	2.64	2.43	2.25	2.71	2.31	**2.61**	2.35	2.43	2.39	2.52	2.53
	$\mathcal{T}$	2.72	2.86	2.31	2.62	2.77	2.52	2.71	2.66	2.51	2.44	2.13	2.01	1.77	2.51	2.20	2.45
	$\mathcal{F}$	2.11	2.21	2.33	N/A	N/A	N/A	N/A	N/A	N/A	2.19	N/A	N/A	2.41	2.32	N/A	2.26
	$\mathcal{S}$	2.51	2.31	2.41	1.81	2.52	2.41	2.12	2.29	2.51	2.13	**2.61**	2.14	2.51	1.87	2.12	2.28
MFN	$\mathcal{W}$	**2.99**	3.01	**2.88**	2.53	2.78	**2.81**	2.77	**2.83**	2.61	2.70	2.56	**2.66**	**2.82**	2.79	2.51	**2.75**
	$\mathcal{H}$	2.81	**3.13**	2.63	2.58	**2.91**	2.77	2.57	2.63	**2.73**	2.50	**2.61**	2.54	2.51	2.69	**2.61**	2.68
	$\mathcal{T}$	2.74	3.03	2.51	**2.64**	2.83	2.51	**2.81**	2.71	2.60	2.48	2.13	2.55	2.19	2.57	2.31	2.57
	$\mathcal{F}$	2.33	2.41	2.34	N/A	N/A	N/A	N/A	N/A	N/A	2.69	N/A	N/A	2.54	2.48	N/A	2.46
	$\mathcal{S}$	2.61	2.44	2.55	2.22	2.61	2.52	2.71	2.31	2.62	2.48	**2.61**	2.31	2.53	2.23	2.13	2.46

- On average, MFN has better performance over MF and NN
- In Mexico, MF has the best accuracy - possibly because the 2013 ILI season in Mexico was late by a few weeks than in usual.

Model level fusion

- Output from models combined based on historical accuracy.

Model level fusion

- Output from models combined based on historical accuracy.
- Model

$$ {}_{C}\mathcal{M}_t = \left[\begin{array}{cccc} {}_1\widehat{P}_t & \dots & {}_C\widehat{P}_t & P_t \end{array} \right] \tag{6} $$

Model level fusion

- Output from models combined based on historical accuracy.
- Model

$$ {}_{C}\mathcal{M}_t = \left[\begin{array}{cccc} {}_{1}\widehat{P}_t & \dots & {}_{C}\widehat{P}_t & P_t \end{array}\right] \tag{6} $$

- Fitting

$$ \begin{split} {}_{C}\widehat{\mathcal{M}}_{i,j} = \mu_i + {}_{C}b_j + {}_{C}U_i^T {}_{C}F_j \\ + {}_{C}F_j |{}_{C}\mathcal{N}(i)|^{-\frac{1}{2}} \textstyle\sum_{k \in {}_{C}N(i)} ({}_{C}\mathcal{M}_{i,k} - \mu_i + {}_{C}b_k) {}_{C}x_k \end{split} \tag{7} $$

Data level fusion

- Feature vector is a tuple over all data set features.

$$\mathcal{X}_t = \langle \mathcal{T}_t, \mathcal{W}_t \rangle$$

- Use MFN to fit the value

Accuracy comparison

Table 2: Comparison of prediction accuracy while combining all data sources and using MFN regression.

Fusion Level	AR	BO	CL	CR	CO	EC	GF	GT	HN	MX	NI	PA	PY	PE	SV	All
Model	**3.12**	**3.22**	3.03	**2.88**	**2.98**	**3.13**	2.87	**2.99**	**2.87**	**3.00**	**2.77**	**2.82**	2.81	**2.92**	2.87	**2.95**
Data	3.01	2.97	**3.13**	2.87	2.86	3.04	**2.91**	2.88	2.72	2.89	2.70	2.60	**2.88**	2.81	**2.92**	2.88

Accuracy comparison

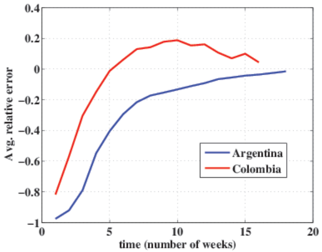
Table 2: Comparison of prediction accuracy while combining all data sources and using MFN regression.

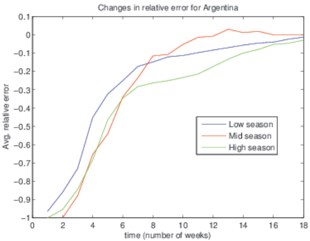
Fusion Level	AR	BO	CL	CR	CO	EC	GF	GT	HN	MX	NI	PA	PY	PE	SV	All
Model	**3.12**	**3.22**	3.03	**2.88**	**2.98**	**3.13**	2.87	**2.99**	**2.87**	**3.00**	**2.77**	**2.82**	2.81	**2.92**	2.87	**2.95**
Data	3.01	2.97	**3.13**	2.87	2.86	3.04	**2.91**	2.88	2.72	2.89	2.70	2.60	**2.88**	2.81	**2.92**	2.88

- On average, model level fusion produces better accuracy than data level fusion.
- Interesting deviations like Chile and El Salvador indicates that a possible strategy could be a mix of data level and model fusion - however complexity of training will increase manifold.

Uncertainty in official estimates

- Can take up to several months to stabilize.





- Average relative error of PAHO count values with respect to stable values. (a) Comparison between Argentina and Colombia (b) Comparison between different seasons for Argentina.

Correcting uncertainty

- Recognize high, low and mid-season months for countries.
- Variable setup

$$\mathcal{P}_A{}^S = \left\{ (1, P_i^{(1)}, \dot{P}_i, N_i^{(1)}), ..., (m, P_i^{(m)}, \dot{P}_i, N_i^{(m)}), ... \right\}$$

- Correction Model

$$\hat{\dot{P}}_i^{(m)} = a_0 + a_1 m + a_2 P_i^{(m)} + a_3 N_i^{(m)} \tag{8}$$

Correcting uncertainty

- Recognize high, low and mid-season months for countries.
- Variable setup

$$\mathcal{P}_A{}^S = \left\{(1, P_i^{(1)}, \dot{P}_i, N_i^{(1)}), ..., (m, P_i^{(m)}, \dot{P}_i, N_i^{(m)}), ...\right\}$$

- Correction Model

$$\hat{\dot{P}}_i^{(m)} = a_0 + a_1 m + a_2 P_i^{(m)} + a_3 N_i^{(m)} \tag{8}$$

Table 3: Comparison of prediction accuracy while using model level fusion on MFN regressors and employing PAHO stabilization.

Correction Method	AR	BO	CL	CR	CO	EC	GF	GT	HN	MX	NI	PA	PY	PE	SV	All
None	3.12	3.22	3.03	2.88	**2.98**	3.13	2.87	2.99	2.87	3.00	2.77	2.82	2.81	**2.92**	2.87	2.95
Weeks Ahead	3.15	**3.24**	3.04	2.87	2.97	3.17	2.87	2.99	2.88	3.05	2.77	2.91	3.02	2.91	**2.88**	2.98
Num. samples	3.20	**3.24**	3.03	2.88	2.96	3.12	2.87	**3.01**	**2.89**	3.12	**2.78**	2.92	3.04	2.91	2.87	2.99
Combined	**3.21**	**3.24**	**3.05**	**2.89**	2.96	**3.19**	**2.88**	3.00	**2.89**	**3.13**	2.77	**2.93**	**3.08**	**2.92**	**2.88**	**3.00**

Investigating importance of each source : Ablation Test

Table 4: Discovering importance of sources in Model level fusion on MFN regressors by ablating one source at a time.

Sources	AR	BO	CL	CR	CO	EC	GF	GT	HN	MX	NI	PA	PY	PE	SV	All
All	**3.21**	**3.24**	**3.05**	**2.89**	**2.96**	**3.19**	2.87	**3.00**	**2.89**	**3.13**	**2.77**	**2.93**	**3.08**	**2.92**	**2.88**	**3.00**
w/o $\mathcal{W}$	2.91	2.99	2.77	2.71	2.61	2.59	2.66	2.69	2.49	2.78	2.62	2.87	2.60	2.43	2.67	2.69
w/o $\mathcal{H}$	3.04	2.85	2.89	2.56	2.81	2.77	2.61	2.75	2.75	2.82	2.57	2.75	2.51	2.87	2.71	2.75
w/o $\mathcal{T}$	2.92	3.14	2.95	2.61	2.72	2.81	2.88	2.79	2.61	2.93	2.74	2.63	2.79	2.74	2.81	2.80
w/o $\mathcal{S}$	3.19	3.11	2.92	2.64	2.69	2.70	**2.89**	2.88	2.78	3.07	2.75	2.91	2.80	2.71	2.86	2.86
w/o $\mathcal{F}$	3.20	3.12	2.88	**2.89**	**2.96**	**3.19**	2.87	**3.00**	2.83	3.02	**2.77**	**2.93**	2.98	2.88	**2.88**	2.96