DARPA PPAML Challenge Problem #7: Flu Spread

Version 1.00, 14. January 2016

Problem Description

Predicting the spread of epidemics through space and time can help government agencies and organizations better prepare and allocate resources. Seasonal flu epidemics have been closely monitored, and many years of historical data have been collected by the medical community. The data collected and aggregated by CDC has been valuable for researchers trying to develop models to forecast the spread of flu epidemics. In addition to the CDC data, there are many other data collected by different entities for various purposes – many of them unrelated to flu epidemics. When those datasets are considered together with the CDC data, they offer the opportunity to significantly improve our ability to assess and forecast flu epidemics both spatially and temporally. Such data include weather data, social network data, vaccination statistics, and flu medication sales. Those data have different characteristics (e.g. percentages for CDC regional ILI rates and flu vaccination, and quantized flu activity levels for CDC state ILI rates) and different spatial and temporal resolution. Aggregating the data into a forecasting model is challenging but, if successful, can provide much improved forecasting accuracy over a longer time horizon than what current approaches based on limited sources of information can accomplish.

Phase 1 Problem

During Phase 1, the goal is to estimate Influenza-like Illness (ILI) rates at a spatial resolution finer than that of the ILI data from CDC. Performers can use all of the datasets listed in the next section, except for the NREVSS dataset (Phase 2 data), to estimate weekly ILI rates in the 48 contiguous states. The spatial resolution of the estimate should be at the county level. The results will be compared to state ILI rates from selected states (Maryland, North Carolina, Rhode Island and Texas) and district ILI rates from 2 states (Mississippi and Tennessee), where each district consists of multiple counties. The development in the first phase also helps to identify important covariates and their contributions to spatio-temporal interpolation and prediction.

The datasets cover the flu seasons 2013-2014 and 2014-2015. Performers can use all or some of the datasets for their development. The public set consists of data from the 2013-2014 flu season. The 2014-2015 flu season data will be used as the private set for evaluation. All of the 2014-2015 Phase 1 data except for the state and district level ILI rates from the selected states are available as input to the models. Fitted models will be evaluated based on their estimated state and district ILI rates against the actual ILI rates of the select states and districts.

Phase 2 Problem

During Phase 2, the goal is to produce estimates ILI rates that are more timely than those published by the CDC data (while maintaining finer spatial resolution finer than the CDC). The ILI data from CDC and states have 1 to 2 weeks of delay. The models developed by the performers will be used to nowcast the weekly state and district ILI rates of the select states. The nowcast results will be compared to the released data from CDC and select states. Performers can also use the NREVSS dataset in this phase, which may provide additional predictive power.

All of the 2013-2015 data are available to performers in the public set. Performers are to develop models for 2-week-ahead nowcasting of the select states' and districts' ILI rates. Input to the model can include all of the data with 2 weeks of delay except for the Twitter data, where the data from the current week can be used. In Phase 2, the evaluation will be conducted on the current 2015-2016 flu season. Performers are evaluated based on their nowcast state and district ILI rates against the actual ILI rates.

Data Description

Main Datasets
CDC Seasonal ILI Rate (HHS Regions)
CDC reports the weekly number of Influenza-like Illness (ILI)* cases collected through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) with roughly 2 weeks' delay. This is considered a highly authoritative source of influenza related information in the medical field. The publicly available dataset contains the percentages of weekly outpatient visits for ILI since 1997. The cases in each reporting period are also divided into 10 Health and Human Services (HHS) regions, where a region covers multiple states, and is partitioned into 5 age groups.
CDC Seasonal Flu Activity Level (States)
CDC also publishes a weekly measure of flu activity for each state. The flu activity is quantized into 10 levels. "The 10 activity levels correspond to the number of standard deviations below, at or above the mean for the current week compared to the mean of the non-influenza weeks. … An activity level of 1 corresponds to values that are below the mean, level 2 corresponds to an ILI percentage less than 1 standard deviation above the mean, level 3 corresponds to ILI more than 1, but less than 2 standard deviations above the mean, and so on, with an activity level of 10 corresponding to ILI 8 or more standard deviations above the mean."
State and County ILI Rates (Selected states and counties)
This dataset contains percentages of ILI cases from Massachusetts, North Carolina, Rhode Island and Texas for the 2012-2013, 2013-2014, and 2014-2015 flu seasons. For Mississippi and Tennessee, percentages of ILI cases are also broken out by districts/regions. Mississippi is divided into 9 districts and Tennessee into 13 regions, where each district/region consists of multiple counties. However, some districts/regions in some weeks have missing ILI data.
Twitter Data
The dataset contains the number of flu related tweets without re-tweets and not from the same user

* "ILI is defined as fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat without a KNOWN cause other than influenza."

within the syndrome elapsed time of 1 week. The flu related tweets are defined as tweets with keywords “flu,” and “influenza.” The locations of the tweets provides observations with finer spatial and temporal resolution than CDC data, but the data are very noisy.
Flu Vaccination Data of Medicare Recipients
As people receive flu vaccines, the percentage of population susceptible to flu is reduced, and even when vaccinated people get the flu, they generally have milder symptoms. This reduces the number of reported ILI cases. Hence, the flu vaccination data may have strong predictive power for future ILI cases. This dataset contains weekly cumulative percentages of Medicare recipients filing flu vaccination claims of each year between 2012 and 2015 for each county in the United States. It is noted that the data covers only Medicare recipients, and the majority of the recipients are age 65 or older.
NREVSS: The National Respiratory and Enteric Virus Surveillance System (Phase 2)
In addition to the ILINet data, CDC also aggregates to HHS regions the weekly percentage and number of respiratory specimens tested positive for influenza and the weekly number of cases for each influenza virus type. Since illness due to Type A and Type B flu virus may peak at different time during the flu season, using this dataset may improve prediction of ILI rates.
Supporting Data
Twitter User Demographics
To account for the demographic differences of Twitter users from the general population, we will use data from a study by Pew Research, that summarizes the demographics of Twitter users by gender, race, education, and income level. The demographic information may help to correct some Twitter sampling bias by normalizing the tweet counts based on the demographics of their corresponding local populations.
US Census
Total population and percentages of population by age group, education and income level of each county in the United States.
County Adjacency Data
The file contains state association of each county and its neighboring counties.

Baseline Spatio-temporal Model

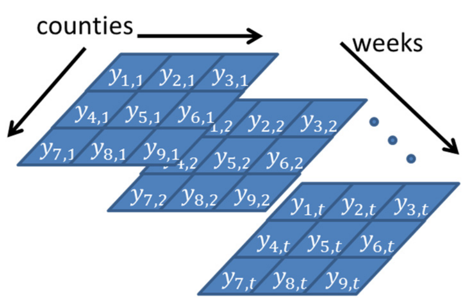
The baseline Bayesian hierarchical model captures various effects commonly present in temporally and geographically referenced data. The baseline model attempts to describe both spatial clustering and temporal correlations. For effects that are caused by known factors, the model also incorporates the fixed-effects term to capture their influence on the observations, in addition to the spatial and temporal random effects. The model has been developed with computational concerns in mind. The use of Gaussian Markov Random Field (GMRF) in modeling the random effects allows this approach to take full advantage of the spatial and temporal correlations while still maintaining manageable memory usage. The specific model structure is below:

County ILI Rate Model	$\log\left(\frac{z_{c,t} + \epsilon_1}{1 - z_{c,t} + \epsilon_1}\right) = \beta^T X_{c,t} + y_{c,t} + n_{c,t}$

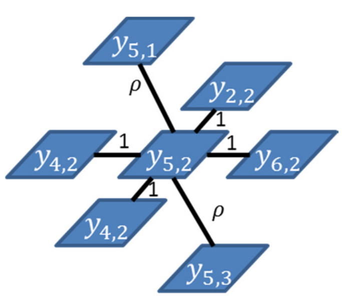
	c: county index; t: week index $\boldsymbol{X_{c,t}}$: covariates for each county $\boldsymbol{c}$ and week $\boldsymbol{t}$ $\boldsymbol{y_{c,t}}$: spatio-temporal correlation of ILI $-\infty < \boldsymbol{y_{c,t}} < +\infty$ (see below) $\boldsymbol{z_{c,t}}$: ILI rate (between 0 and 1) of county c in week t $\boldsymbol{n_{c,t}}$: zero-mean Gaussian noise (see below) $\boldsymbol{\epsilon_1 = 0.0001}$: a small number to ensure numerical stability
Aggregate Observations	$z_{R_i,t} = \sum_{c \in R_i} \left(\frac{N_c}{N_{R_i}}\right) z_{c,t}$ R_i : set of counties in area i; the area can be a HHS Region, a state or a district in a state $z_{R_i,t}$: reported ILI rate of HHS Region i in week t N_c : population of county c N_{R_i} : population of region i
Spatio-temporal correlation	$p(Y) \propto \exp\left(-\frac{1}{2}\tau_1 Y^T (D_w - W)\, Y\right)$ $Y = [y_{1,0} \quad y_{2,0} \quad \ldots \quad y_{1,1} \quad y_{2,1} \quad \ldots \quad y_{C,T-1} \quad y_{C,T}]^T$ C: total number of counties; T: total number of weeks D_w: diagonal matrix; W: sparse symmetric matrix The elements of W are defined as $\begin{cases} w_{(c,t)(c,j)} = \rho & \text{where } j = t-1 \text{ or } t+1, \\ w_{(c,t)(i,t)} = 1 & \text{if } i \text{ is a neighboring county of } c, \\ w_{(c,t)(i,j)} = 0 & \text{otherwise,} \end{cases}$ where the subscript (c, t) (i, j) denotes the row and column indices of the element of W that corresponds to $y_{c,t}$ and $y_{i,j}$ in Y. It is noted that W is a very large but very sparse matrix. However, the only non-zero entries of W are those whose row and column indices correspond to pairs of elements in Y that represent the effects of neighboring counties from the same week or of the same county from consecutive weeks. Hence, it is recommended using a sparse matrix formulation for W and similarly for D_w $(D_w)_{(c,t)(c,t)} = \Sigma_{(i,j)} w_{(c,t)(i,j)}$ τ_1 : precision (inverse variance) that controls the spatio-temporal correlations (smoothness) of ILI rates
Covariates	$X_{c,t} = \left[\log\left(\frac{S_{c,t} + \epsilon_2}{\tilde{N}_c}\right), \quad \log\left(\frac{V_{c,t} + \epsilon_3}{1 - V_{c,t} + \epsilon_3}\right)\right]^{\boldsymbol{T}}$ $S_{c,t}$: number of flu-related tweets from county c in week t. $V_{c,t}$: cumulative percentage of Medicare recipients filing flu vaccination claims from county c in

	week t. $\widetilde{N}_c = \Sigma_g N_{c,g} U_g$: Twitter user demographics adjusted population of county c $N_{c,g}$: population of county c belonging to age group g. U_g : percentage of Twitter users belonging to age group g. $\epsilon_2 = 0.1,\ \epsilon_3 = 0.001$
County heterogeneity	$n_{c,t}\ \ \sim^{iid}\ N(0, 1/\tau_2\ )$
Hyperprior	$\beta_1\ \sim N(0,10)$, $\beta_2\ \sim N(0,10)$ $\tau_1 \sim G(3, 0.1)$, Gamma distribution (α (shape), β (rate) definition) $\tau_2 \sim G(10, 0.1)$ $\rho\ \sim$ G(1.05, 0.5)

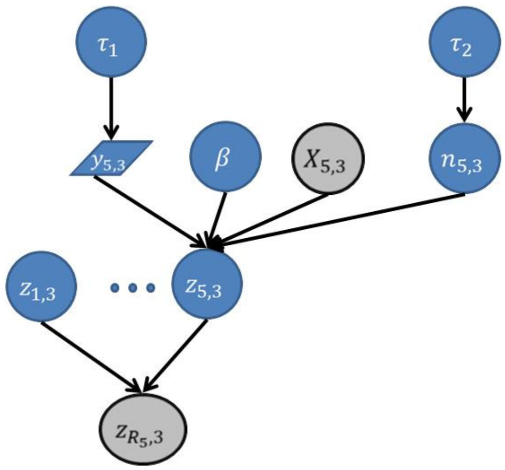
We can visualize the model as follows where we assume that the counties are located on a grid for illustration purpose. The GMRF defines a time series of latent prevalence maps:



Each cell is connected to its neighbors in time and space. The adjacency matrix W contains a 1 for spatial neighbors and ρ for temporal neighbors.



The latent prevalence is combined with covariates and a noise term to generate the observations:



Grey variables are observed at training time. Only X is observed at prediction time. Note that typically several individual county prevalences ($z_{1,3}, \dots, z_{5,3}$) are aggregated to produce the observed value.

Queries

1. **Reconstruction.** Given weekly $S_{c,t}$ (number of flu related tweets) and $V_{c,t}$ (cumulative percentage of Medicare flu vaccination claims) of every county and weekly $z_{R_i,t}$ (ILI rates) of every HHS region, output a marginal MAP and MAP estimates for weekly ILI rates $z_{R_D,t}$ of individual counties in the selected states (Tennessee, Mississippi, Massachusetts, North Carolina, Rhode Island and Texas).
2. **Prediction.** Given weekly $S_{c,t}$ (number of flu related tweets) and $V_{c,t}$ (cumulative percentage of Medicare flu vaccination claims) of every county for weeks $t = 1, \dots, u$ and weekly $z_{R_i,t}$ (ILI rates) of every HHS region for weeks $t = 1, \dots, u - 1$, output a marginal MAP and MAP estimates for weekly ILI rates $z_{R_D,u}$ of individual counties in the selected states (Tennessee, Mississippi, Massachusetts, North Carolina, Rhode Island and Texas).

Evaluation Metrics

1. **Population-adjusted Squared Error.** Let $\hat{z}_{c,t}$ be the predicted prevalence for county c at time t; let R_i be a region that contains county c; let $z_{R_i,t}$ be the observed prevalence in region R_i; and let N_c be the population in count c. A region in this case is the smallest geographical area where ILI truth data is available. Define $\hat{z}_{R_i,t}$ to be the aggregated prevalence (weighted by relative population):

$$\hat{z}_{R_i,t} = \frac{\sum_{c \in R_i} N_c \hat{z}_{c,t}}{\sum_{c \in R_i} N_c}.$$

Then the loss is

$$\sum_i N_{R_i} \sum_t \left(\hat{z}_{R_i,t} - z_{R_i,t}\right)^2$$

2. **Start and Max of the epidemic**. Let $\overline{M}_{R_i}$ be the week in which the highest prevalence was observed during the year. Let S_{R_i} be the week when the prevalence first exceeded 5%. We will compute the absolute difference between the observed and predicted values of these quantities.

Data Files

The following data files are provided in the “CP7-FluSpread” directory of the MIDAS data set repository to assist you in the development of a solution:

Synthetic data to ***debug*** your solution:

- **`Flu_Vacc_Syn.json`**
- **`Output_Syn.json`**

`Output_Syn.json` is the synthetic truth output from the baseline model when the synthetic data **`Flu_Vacc_Syn.json`** is used as input and no random effect terms are used.

Actual data to ***train*** your solution:

- **`Flu_Vacc_Tweet_TRAIN.json`**
- **`Flu_ILI_TRAIN.csv`**

Actual data to ***test*** your solution:

- **`Flu_Vacc_Tweet_TEST.json`**
- **`Flu_ILI_TEST.csv`**

Provided supporting data:

`StateInfo.json` contains the names and FIPS codes of the counties in each state

`Region2CountyMap.json` contains the mappings from individual counties to HHS Regions, selected states, and districts within selected states. For example, the object “HHS Region 1” contains all counties (FIPS code and name) that belong to HHS Region 1. Similarly, the object "TN District 7" contains all the counties in District 7 of Tennessee. The object names (e.g. "TN District 7") are the same as the headers in Flu_ILI_TRAIN.csv for ease of reference.

county_adjacency_lower48.json contains adjacent counties of each county in the lower 48 states. It helps create the adjacency matrix in the baseline model.

PI_SocialMediaUpdate20144.pdf: Pew Research paper on Twitter demographics (page 6)

Input and Output Data Formats

Input 1 (Flu_Vacc_Syn.json, Flu_Vacc_Tweet_{TRAIN | TEST}.json): $S_{c,t}$ (number of flu-related tweets) **and $V_{c,t}$ (**cumulative percentage of Medicare recipients) stored in a **JSON** file. The name of each object is the 5-digit FIPS code of a county (or equivalent), which uniquely denotes a county (e.g. 01001 is Autauga County, AL), and its value is relevant information associated with the county including $S_{c,t}$ (weekly number of flu-related tweets), $V_{c,t}$ (weekly cumulative percentage of Medicare recipients filing flu vaccination claims) and demographics.


```
"01001": {
 "Population, 2010": 54571,
 "Per capita money income in past 12 months (2013 dollars), 2009-2013": 24571,
 "Persons under 18 years, percent, 2013": 25.4,
 "Vaccination percentage %": {
  "08/10/2013": 0.0,
  "08/17/2013": 0.0,
  "08/24/2013": 0.0,
  "08/31/2013": 1.36001554,
  "09/07/2013": 3.53604041,
  "09/14/2013": 6.25607148,
  "09/21/2013": 10.43326207,
  "09/28/2013": 15.44589079,
  "10/05/2013": 21.391101600000002,
  "10/12/2013": 26.85059253,
        :
        :
  "07/12/2014": 46.53196028,
  "07/19/2014": 46.53196028,
  "07/26/2014": 46.53196028,
  "08/02/2014": 46.53196028
 },
 "Name": "Autauga County, AL",
 "High school graduate or higher, percent of persons age 25+, 2009-2013": 85.6,
 "Population, 2014 estimate": 55395,
 "Persons 65 years and over, percent, 2013": 13.5,
 "Persons under 5 years, percent, 2013": 6.1,
 "Median household income, 2009-2013": 53682,
 "Population, 2013 estimate": 55136,
 "FIPS State and County code": 1001,
 "Bachelor's degree or higher, percent of persons age 25+, 2009-2013": 20.9,
 "No. of Tweets": {
  "08/10/2013": 1,
  "08/17/2013": 0,
  "08/24/2013": 1,
        :
        :
 }
```

Input 2 (Flu_ILI_{TRAIN | TEST}.csv): Weekly $z_{R_i,t}$ (ILI rates) of HHS Regions, states and districts in CSV. If no data is available, NaN is given. An example of Flu_ILI_Train.csv is illustrated below.

Note: Flu_ILI_TEST.csv differs from Flu_ILI_TRAIN.csv in that Flu_ILI_TEST.csv only contains ILI rates of HHS Regions, while Flu_ILI_TRAIN.csv contains ILI rates from selected states and districts in some states in addition to the HHS Regions.

Year	Week #	Ending	HHS Region 1	HHS Region 2	HHS Region 3		HHS Region 9	HHS Region 10	MA	MS	NC	
2013	31	8/3/2013	0.47%	0.97%	0.76%		0.94%	0.12%	0.48%	NaN	NaN	
2013	32	8/10/2013	0.53%	0.92%	0.65%		1.00%	0.22%	0.59%	NaN	NaN	
2013	33	8/17/2013	0.45%	0.79%	0.75%		0.98%	0.17%	0.45%	NaN	NaN	
2013	34	8/24/2013	0.45%	0.86%	0.72%		0.85%	0.16%	0.53%	NaN	NaN	
2013	35	8/31/2013	0.41%	0.82%	0.78%		1.19%	0.16%	0.38%	NaN	NaN	
2013	36	9/7/2013	0.35%	0.81%	0.95%		1.30%	0.16%	0.26%	NaN	NaN	
2013	37	9/14/2013	0.40%	0.86%	1.13%		1.32%	0.26%	0.43%	NaN	NaN	
2013	38	9/21/2013	0.47%	1.23%	1.22%		1.35%	0.32%	0.53%	NaN	NaN	
2013	39	9/28/2013	0.61%	1.42%	1.23%		1.61%	0.34%	0.50%	2.60%	NaN	
2013	40	10/5/2013	0.62%	1.43%	1.09%		1.11%	0.47%	0.62%	3.10%	0.35%	
2013	41	10/12/2013	0.64%	1.35%	1.17%	....	1.34%	0.45%	0.70%	2.90%	0.86%	....
2013	42	10/19/2013	0.73%	1.46%	1.23%		1.33%	0.55%	0.84%	3.60%	0.74%	
2013	43	10/26/2013	0.69%	1.61%	1.33%		1.28%	0.50%	0.71%	4.10%	0.73%	
2013	44	11/2/2013	0.72%	1.69%	1.33%		1.46%	0.53%	0.75%	4.70%	0.82%	
2013	45	11/9/2013	0.78%	1.75%	1.34%		1.68%	0.40%	0.79%	4.80%	0.89%	
2013	46	11/16/2013	0.79%	1.76%	1.44%		1.62%	0.53%	0.86%	5.30%	0.99%	
2013	47	11/23/2013	0.71%	1.91%	1.46%		1.66%	0.48%	0.81%	5.00%	0.87%	
2013	48	11/30/2013	0.93%	2.25%	1.76%		1.95%	0.66%	1.05%	6.20%	1.45%	
2013	49	12/7/2013	0.82%	2.30%	1.77%		1.84%	0.64%	0.84%	6.00%	1.23%	
2013	50	12/14/2013	1.03%	2.32%	1.88%		2.09%	0.92%	1.24%	6.60%	1.47%	
2013	51	12/21/2013	1.14%	2.63%	2.30%		2.57%	1.81%	1.31%	7.70%	2.07%	
2013	52	12/28/2013	1.74%	3.37%	3.30%		3.70%	4.73%	1.88%	8.00%	4.74%	
2014	1	1/4/2014	1.74%	3.50%	3.74%		3.74%	4.99%	1.94%	7.50%	3.91%	
2014	2	1/11/2014	1.42%	2.86%	3.72%		3.46%	3.90%	1.33%	6.30%	2.70%	
2014	3	1/18/2014	1.74%	3.12%	4.15%		3.58%	2.96%	1.82%	5.50%	3.25%	
2014	4	1/25/2014	1.88%	3.03%	4.11%		3.85%	2.40%	1.74%	5.10%	3.18%	
2014	5	2/1/2014	1.82%	2.82%	3.37%		3.44%	1.88%	1.58%	4.90%	2.90%	
2014	6	2/8/2014	2.09%	2.81%	3.03%		3.10%	1.65%	2.00%	4.20%	1.76%	
2014	7	2/15/2014	1.68%	2.54%	2.81%		2.72%	1.27%	1.65%	3.80%	2.06%	
2014	8	2/22/2014	1.54%	2.27%	2.39%		2.50%	1.24%	1.46%	3.60%	1.43%	

Output 1: Weekly $z_{R_i,t}$ (ILI rates) of individual counties in a JSON file. The name of each object is the 5-digit FIPS code of a county, and its value contains its weekly ILI rates and in the time frame of interest:


```
{
 "01001": {
  "ILI percentage %": {
   "08/10/2013": 0.201,
   "08/17/2013": 0.198,
   "08/24/2013": 0.203,
   "08/31/2013": 1.360,
   "09/07/2013": 3.536,
   "09/14/2013": 4.256,
       :
       :
   "08/15/2015": 0.101,
   "08/22/2015": 0.102,
   "08/29/2015": 0.664,
   "09/05/2015": 1.836,
   "09/12/2015": 2.798,
   "09/19/2015": 2.798,
   "09/26/2015": 2.798
  },
  "Name": "Autauga County, AL",
 },
 "01003": {
  "ILI percentage %": {
       :
       :
 },
       :
}
```

Provided Supporting Data

StateInfo.json: a JSON file that provides the names and FIPS codes of the counties in each state.


```
{
 "AK": {
  "FIPS": "02000",
  "Name": "ALASKA",
  "Counties": {
   "02180": "Nome Census Area",
   "02185": "North Slope Borough",
   "02188": "Northwest Arctic Borough",
   "02275": "Wrangell City and Borough",
      :
      :
   "02016": "Aleutians West Census Area",
   "02240": "Southeast Fairbanks Census Area",
   "02013": "Aleutians East Borough",
   "02282": "Yakutat City and Borough",
   "02070": "Dillingham Census Area"
  }
 },
 "AL": {
  "FIPS": "01000",
  "Name": "ALABAMA",
  "Counties": {
   "01069": "Houston County",
   "01079": "Lawrence County",
   "01011": "Bullock County",
   "01045": "Dale County",
   "01043": "Cullman County",
   "01027": "Clay County",
      :
      :
  }
 },
  :
}
```

Region2CountyMap.json: contains the mappings from individual counties to HHS Regions, selected states, and districts within selected states.


```
{
 "HHS Region 1": {
  "09007": "Middlesex County",
  "09005": "Litchfield County",
  "44009": "Washington County",
  "09003": "Hartford County",
  "09001": "Fairfield County",
  "44005": "Newport County",
  "44007": "Providence County",
  "44001": "Bristol County",
  "44003": "Kent County",
  "09009": "New Haven County",
  "25019": "Nantucket County",
  "25013": "Hampden County",
  "25011": "Franklin County",
  "25017": "Middlesex County",
  "25015": "Hampshire County",
      :

      :

  "33001": "Belknap County",
  "23023": "Sagadahoc County",
  "33007": "Coos County",
  "23025": "Somerset County",
  "33005": "Cheshire County",
  "23027": "Waldo County"
 },
 "HHS Region 10": {
  "41019": "Douglas County",
  "41055": "Sherman County",
  "16079": "Shoshone County",
  "41051": "Multnomah County",
  "41053": "Polk County",
  "41011": "Coos County",
  "53033": "King County",
  "16071": "Oneida County",
  "16015": "Boise County",
      :
```

county_adjacency_lower48.json: contains adjacent counties of each county in the lower 48 states. The 5-digit numbers are the FIPS codes of the corresponding counties.


```
{
  "Autauga County, AL": [
    "AL",
    "01001",
    {
      "Chilton County, AL": "01021",
      "Lowndes County, AL": "01085",
      "Dallas County, AL": "01047",
      "Montgomery County, AL": "01101",
      "Autauga County, AL": "01001",
      "Elmore County, AL": "01051"
    }
  ],
  "Baldwin County, AL": [
    "AL",
    "01003",
    {
      "Escambia County, FL": "12033",
      "Washington County, AL": "01129",
      "Baldwin County, AL": "01003",
      "Clarke County, AL": "01025",
      "Monroe County, AL": "01099",
      "Mobile County, AL": "01097",
      "Escambia County, AL": "01053"
    }
  ],
  "Barbour County, AL": [
    "AL",
    "01005",
    {
      "Clay County, GA": "13061",
      "Stewart County, GA": "13259",
      "Henry County, AL": "01067",
      "Barbour County, AL": "01005",
      "Bullock County, AL": "01011",
      "Pike County, AL": "01109",
      "Dale County, AL": "01045",
      "Quitman County, GA": "13239",
      "Russell County, AL": "01113"
    }
      :

      :
```

PI_SocialMediaUpdate20144.pdf: Pew Research paper on Twitter demographics (page 6)

Twitter — 23% of adult internet users/19% of entire adult population

Some 23% of online adults currently use Twitter, a statistically significant increase compared with the 18% who did so in August 2013. Twitter is particularly popular among those under 50 and the college-educated. Compared with late 2013, the service has seen significant increases among a number of demographic groups: men, whites, those ages 65 and older, those who live in households with an annual household income of $50,000 or more, college graduates, and urbanites.

Twitter users

Among online adults, the % who use Twitter

	2013	2014
All internet users	*18%*	*23%**
Men	17	24*
Women	18	21
White, Non-Hispanic	16	21 *
Black, Non-Hispanic	29	27
Hispanic	16	25
18-29	31	37
30-49	19	25
50-64	9	12
65+	5	10*
High school grad or less	17	16
Some college	18	24
College+ (n= 685)	18	30*
Less than $30,000/yr	17	20
$30,000-$49,999	18	21
$50,000-$74,999	15	27*
$75,000+	19	27*
Urban	18	25*
Suburban	19	23
Rural	11	17

Source: Pew Research Center's Internet Project September Combined Omnibus Survey, September 11-14 & September 18-21, 2014. N=1,597 internet users ages 18+. The margin of error for all internet users is +/- 2.9 percentage points. 2013 data from Pew Internet August Tracking Survey, August 07 – September 16, 2013, n= 1,445 internet users ages 18+.

Note: Percentages marked with an asterisk (*) represent a significant change from 2013. Results are significant at the 95% confidence level using an independent z-test.

PEW RESEARCH CENTER