DARPA PPAML Challenge Problem #7:

Flu Spread

**Version 1.04, 3 May 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. Datasets include social network data and vaccination statistics. 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 (Reconstruction)

During Phase 1, the goal is to fuse multiple data sources to reconstruct 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 a set of “Evaluation Regions” consisting of state-level ILI rates from selected states (Massachusetts, North Carolina, Rhode Island and Texas) and district-level ILI rates from 2 states (Mississippi and Tennessee), where each district consists of multiple counties.

The datasets cover the flu seasons 2013-2014 and 2014-2015. For model development and prediction, performers will have access to all data from both years with the exception of the Evaluation Region data for 2014-2015. 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 (Nowcasting)

During Phase 2, the goal is to produce estimates ILI rates that are more timely than those published by the CDC (while maintaining spatial resolution finer than the CDC). The ILI data from CDC and the Evaluation Region states are released after a delay of 1 to 2 weeks. The goal of Phase 2 is to predict ILI rates in week using all data from previous weeks . This will include the CDC ILI rates from week and the Twitter data from week . All data from 2013-2014 and 2014-2015 will be available for model development and training. Performers can also use the NREVSS dataset in this phase, which may provide additional predictive power. Evaluation will be performed on data collected during the current 2015-2016 flu season.

# Data Description

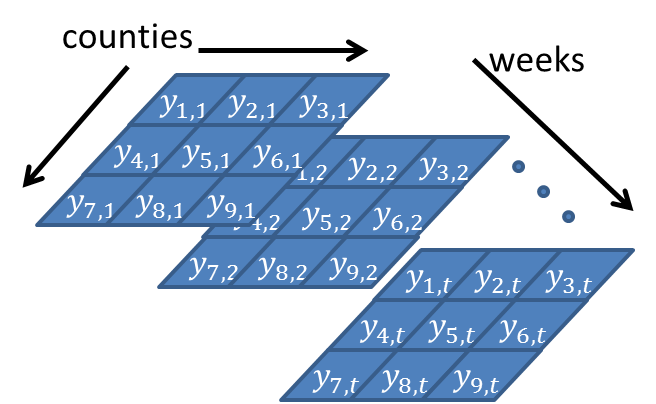
|  |
| --- |
| Main Datasets |
| CDC Seasonal ILI Rate (HHS Regions) |
| CDC reports the weekly number of Influenza-like Illness (ILI)[[1]](#footnote-1) cases collected through the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) with roughly 2 weeks’ delay (<http://www.cdc.gov/flu/weekly/fluactivitysurv.htm>). 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 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 provide 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 these recipients are age 65 or older. |
| NREVSS: The National Respiratory and Enteric Virus Surveillance System (Phase 2 only) |
| 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 times 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 (<http://www.pewinternet.org/2015/01/09/social-media-update-2014/>), 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. |

# Reference Spatio-temporal Model

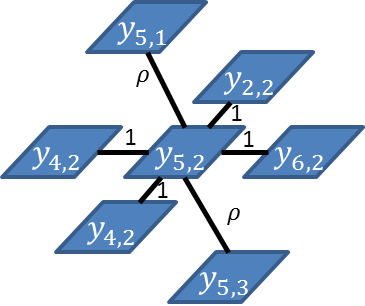
The reference Bayesian hierarchical model captures various effects commonly present in temporally and geographically referenced data. The reference model attempts to describe both spatial clustering and temporal correlations. For effects that are caused by known factors, the model also incorporates a 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 | *c*: county index; *t*: week index  : covariates for each county and week  : spatio-temporal correlation of ILI (see below)  : ILI rate (between 0 and 1) of county *c* in week *t*  : zero-mean Gaussian noise (see below)  : a small number to ensure numerical stability |
| Aggregate Observations | : set of counties in area ; the area can be a HHS Region, a state or a district in a state  : reported ILI rate of HHS Region *i* in week *t*  : population of county c  : population of region *i* |
| Spatio-temporal correlation | The elements of *W* are defined as  where the subscript (*c, t*) (*i, j*) denotes the row and column indices of the element of *W* that corresponds to and in *Y*. It is noted that *W* is a very large but very sparse matrix. 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 to use a sparse matrix representation for *W* and similarly for *Dw*  : precision (inverse variance) that controls the spatio-temporal correlations (smoothness) of ILI rates |
| Covariates | : number of flu-related tweets from county *c* in week *t*.  : cumulative percentage of Medicare recipients filing flu vaccination claims from county *c* in week *t*.  : Twitter user demographics adjusted population of county *c*  : population of county c belonging to age group *g*.  : percentage of Twitter users belonging to age group *g*. |
| County heterogeneity |  |
| Hyperprior | ~ *N*(0,10), ~ *N*(0,10)  ~ *G*(3, 0.1), Gamma distribution ( (shape),  (rate) definition)  ~ *G*(10, 0.1)  ~ G(1.05, 0.5) |
|  |  |

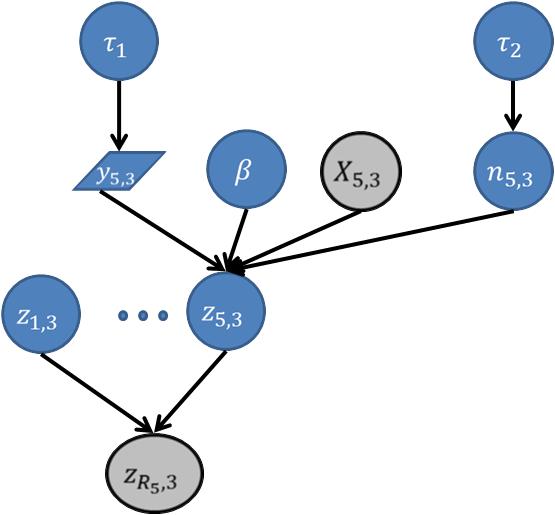
We can visualize the model as follows where we assume that the counties are located on a grid for illustration purpose. The GMRF defines a time series of latent prevalence maps:



Each cell is connected to its neighbors in time and space. The adjacency matrix W contains a 1 for spatial neighbors and 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 is observed at prediction time. Note that typically several individual county prevalences () are aggregated to produce the observed value.

# Queries

1. **Reconstruction.** Given weekly (number of flu related tweets) and (cumulative percentage of Medicare flu vaccination claims) of every county and weekly (ILI rates) of every HHS region, output MAP and/or marginal MAP estimates for weekly ILI rates of individual counties in the selected states (Tennessee, Mississippi, Massachusetts, North Carolina, Rhode Island and Texas).
2. **Nowcasting.** Given weekly (number of flu related tweets) and (cumulative percentage of Medicare flu vaccination claims) of every county for weeks and weekly (ILI rates) of every HHS region for weeks , output MAP and/or marginal MAP estimates for weekly ILI rates of individual counties in the selected states (Tennessee, Mississippi, Massachusetts, North Carolina, Rhode Island and Texas).

# Evaluation Metrics

1. **Population-adjusted Squared Error.** Let be the predicted prevalence for county at time ; let be a region that contains county ; let be the observed prevalence in region ; and let be the population in count . A region in this case is the smallest geographical area where ILI truth data is available. Define to be the aggregated prevalence (weighted by relative population):

Then the loss is

1. **Start and Max of the epidemic**. Let be the week in which the highest prevalence was observed during the year. Let be the week when the prevalence first exceeded 5%. We will compute the absolute difference between the observed and predicted values of these quantities.

# Provided Data Files

The data files for development and testing are found in 3 separate directories – Full/, Middle/, Small/.

The Small/, Middle/ and Full/ directories contain similar sets of data files except that they cover increasingly larger geographical areas. The files in the Small/ directory contain ILI data relevant to Mississippi only. The files in the Middle/ directory contain relevant ILI data for HHS Region 4, which consists of Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee. The Full/ directory covers the lower 48 states. Each directory contains the following main input files in addition to other supporting data, and their formats are defined in the next section.

**Flu\_ILI.csv**: training data and testing data that cover respectively the 2013-2014 and 2014-2015 flu seasons. If no data is available for a particular week and geographical region, then the corresponding cell is labeled ‘NaN’. For the training year (2013-2014), the unavailable data are results of missing information from local health agencies. For the testing year (2014-2015), the NaN cells are mainly due to withheld truth data for evaluation purposes.

**states.txt:** specifying the states of interest for reconstructing and nowcasting county ILI rates. The states are listed in 2-letter state abbreviations. If the file is empty, then the counties in all lower 48 states are of interest. For the Small/ problem, only the counties in Mississippi are of interest. For the Middle/ problem, the 8 states in HHS Region 4 are of interest. For the Full/ problem, all lower 48 states are of interest.

In addition to the above 2 main input files, other relevant data files such as county demographics, geographics and tweet counts are also provided in the directories. The contents of those files are described below, and their formats are given in the section Provided Supporting Data later.

**Flu\_Vacc\_Tweet\_{TRAIN | TEST}.json:** contains weekly flu-related tweet counts, cumulative percentages of Medicare recipients filing flu vaccination claims and demographic information of each county in the lower 48 states. The flu-related tweets are defined as those tweets that contain “flu” or “influenza.” Flu\_Vacc\_Tweet\_TRAIN.json and Flu\_Vacc\_Tweet\_TEST.json are respectively for the 2013-2014 and the 2014-2015 flu seasons.

**Region2CountyMap.json** contains the mappings from HHS Regions, selected states, and districts within selected states to the counties that compose those areas. 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 reference model. Adapted from <https://www.census.gov/geo/reference/county-adjacency.html>

**PI\_SocialMediaUpdate20144.pdf:** Pew Research paper on Twitter demographics (page 6)

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

# Input and Output Data Formats

**Input: Flu\_ILI.csv**

Weekly(ILI rates) of HHS Regions, states and districts in CSV. If no data is available, NaN is given. An example of Flu\_ILI.csv is illustrated below. The first, second and third columns are always, respectively, years, HHS weeks and the dates on which the HHS weeks end; the rest of the columns are the weekly ILI rates for different regions/states/districts. The number of columns varies depending on the problems (small, median or full). The header of each column specifies the region/state/district in which the corresponding ILI data were collected. The counties that compose the region/state/district listed in the headers can be found in the file **Region2CountyMap.json**



**Input: states.txt**

The file specifies the states of interest for reconstructing and nowcasting county ILI rates. Each line in the file lists a state of interest in its 2-letter state abbreviation. If the file is empty, then all lower 48 states are of interest. Hence, for the Full/ problem, the file is empty since all of the counties in the lower 48 states are of interest. The purpose of this file is to allow participants to focus models and solutions only on a subset of states (MS for Small/ and 8 HHS Region 4 states for Middle/) instead of the whole country.

|  |  |
| --- | --- |
| Small/states.txt | Middle/states.txt |
| MS | AL  FL  GA  KY  MS  NC  SC  TN |

**Output: CountyWeeklyILI.json**

Weekly (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 in the time frame of interest. The counties of interest contained in the file should cover all of the counties that compose the geographical area listed in the input file Flu\_ILI.csv. In other words, for the Small/ data directory, the output file should contain all of the counties in Mississippi; for the Middle/ data directory, all of the counties in HHS Region 4; for the Full/ data directory, all of the counties in the lower 48 states.



# Provided Supporting Data Formats

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



**Flu\_Vacc\_Tweet\_{TRAIN | TEST}.json: (**number of flu-related tweets) **and (**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 (weekly number of flu-related tweets)**,**  (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,

:

:

}

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

{

"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",

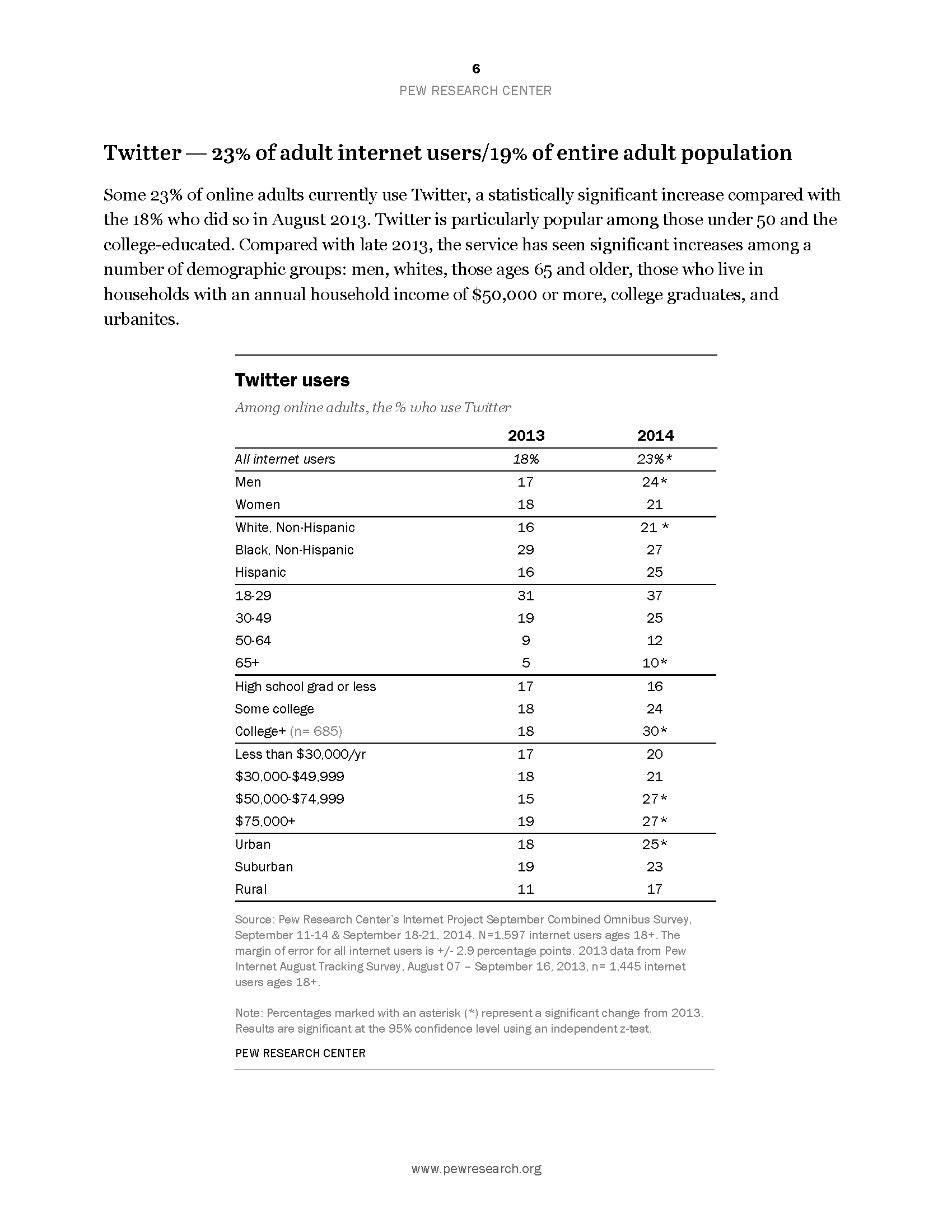
"Russell County, AL": "01113"

}

**:**

**:**

**PI\_SocialMediaUpdate20144.pdf:** Pew Research paper (<http://www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate20144.pdf>) on Twitter demographics (page 6)



1. “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.” [↑](#footnote-ref-1)