Data Analytics – paper 1 February 13,2014
Alex V Flinsch

**Introduction**
I will be working with data from the WSPRnet (Weak Signal Propagation Reporting) project, which may be found at http://wsprnet.org. This data involves high frequency (HF) radio wave propagation at extremely low power levels. Other than a single blog entry from the site maintainer about an attempt to process some of the data in R, in order to generate some usage statistics for the site, I have not seen anything else about analyzing the data.

My main interest is personal, as I have seen several of my signals reported from locations at times/solar conditions where the traditional propagation models show the path as not being possible (although stranger things have happened on HF -- the propagation models are far from perfect). My gut feeling is that the propagation paths actually open/close several hours before/after than what the traditional models predict, or that the paths are always open. Another possibility is that the traditional model assumes a single bounce off the ionosphere between the transmitting and receiving sites, and the heavy error correction and low symbol rate in WSPR allow the decoding of these signals, when the propagation path consists of multiple hops.

WSPRnet began collecting data in July of 2008, and has seen a steady increase in the number of transmitting and reporting stations since then. Currently the WSPR project is collecting around 5 million data entries per month, from amateur radio operators worldwide.

**Available data**
The WSPR software is fundamentally different from other digital amateur modes. Instead of being used for two way communications, it is used in a one way beaconing mode, where signal reports are sent via the internet to a centralized location. The transmitting station selects a percentage of time that they will be transmit for, the remaining time is spent listening for, and reporting on other signals. A 2 minute long signal is sent which contains the following information:

- Call sign[1] of the transmitting station
- Maidenhead grid square[2] in either 4 or 6 characters
- Transmitting power level[3]

Note that this is not very much data, the actual data is encoded and sent twice during each transmission, there is a large amount of forward error correction applied.

Once the signal is received, the reporting station computes and adds the following information:

- Timestamp of the 2 minute starting block in Unix time() format
- Reporting station id (usually an amateur radio call sign)
- Received frequency (sum of the fixed transmission frequency and frequency of the modulating signal)
- Reporting station Maidenhead grid square
- Signal to Noise ratio
- Apparent drift of the transmitted signal if any
- Version number of the WSPR software used at the reception site

After appending the reception data to the transmitted data, the reporting station then forwards to wspnet.org.

When the reported data is received at wsprnet.org, a few additional data fields are added

- Unique spot id used for a database key
- Computed distance between the transmitting and reporting stations
- Compass heading between transmitting and reporting stations
- Amateur radio band used[4]

The data is then formatted into CSV files, and is available in compressed monthly files (current month has data thru the previous day).

In order to correlate the WSPR data to propagation predictions, I will need the daily solar index, flux , and sunspot numbers available from http://www.swpc.noaa.gov/ftpmenu/warehouse.html . Solar flare data is published daily, but the flux, index and sunspot numbers are not available here until the end of a calendar year. The solar index and sunspot numbers are available elsewhere on a realtime basis, but this is the only place where I can find a summary version.

**Previous Studies**

As stated earlier, I was unable to find any studies using WSPRnet data to show the possibility of propagation paths between 2 points opening earlier and closing later than the traditional models predict.

My study will attempt show that 'impossible' paths are possible (and somewhat common) when the traditional models predict otherwise. I would need to base this on the date and time of the transmitted signal, locations of the transmitting and reporting station, frequency used, and solar conditions on the day. At the worst, I will be able to show that greater distances are obtained on days with good (high flux, high sunspot number) days, compared to poorer solar conditions.

**Data Gathering and Preparation**

The WSPR data is available in convenient comma delimited monthly files, which can easily be imported into database or most other software (at file sizes of over 5 million records, they are too large to import into any version of Excel that I know of). On the other hand, the solar condition data is available in a less friendly fixed format plain text file, with around a dozen non-data records at the start of the file. Fortunately these records are identified by either a ':' or a '#' character in the first column, and can easily be eliminated. After stripping these header records, it should be a simple matter to import them into the analysis software.

Since the solar data is only easily available in yearly files, I will be limited to using data from 2013 and earlier. I am inclined to use the data from either 2010 or 2011, as there were quite a few days in those years with zero sunspot counts.

While the WSPR data is in a convenient format, it does contain some data, which is unnecessary for this study. Solar conditions generally do not have much impact on frequencies 30 MHz, and propagation is mostly line of sight. These signal reports will be removed, as there will be little or no correlation to the solar data. I would also be inclined to eliminate any reports below 3 MHz. While solar conditions to have some effect below 3 MHz. The primary propagation mode in these cases would be groundwave rather than skywave. Additionally two of the bands reported by WSPR below 3 MHz are unlicensed experimental bands in the US, and have little interest to me. I will also be deleting any data where the calculated distance is less than 150 km, these are primarily groundwave signals and if left in the data, would only create unnecessary noise. Many fields from the data could also be eliminated. About the only data that would really be required, would be the timestamp, grid locators, and the frequency band in use.

So far, I have already created a MySQL database to enable easy selection of the WSPR data, by date and band. I have also successfully loaded several months of data. I was also able to delete the data points that will be unused.

**Notes**

1 – Call Sign – license assigned by the communications regulatory body of the country where the station is located.

2 – Maidenhead grid square, a shorthand method of denoting latitude & longitude consisting of alternating letter/number pairs, mostly used by amateur radio operators. Four position locator squares represent an accuracy of 1° of latitude by 2° of longitude, six character locators improve the accuracy to 2.5' of latitude by 5' of longitude.

3 – Power level as specified in dBm – decibel increase relative to 1mW. Values are mostly positive, but can go negative for stations transmitting at < 1mW.

4 – Amateur radio band, in most cases this field contains the integer portion of the received frequency, for the 470 kHz experimental band, it is specified as 0, for the 35 kHz experimental band it is specified as -1.