Alex V Flinsch
Data Analytics – Paper 2

Introduction

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 will be using data from both January 2013, and January 2009. These years and months were chosen for two reasons. The wspr data from January 2013 was the most recent available when I started this project, and the corresponding month in 2009, had quite a few days with zero sunspots, hopefully this will allow for easy visual interpretation of any changes in the data. Additionally I chose to concentrate on the data available for a single band (30 meters, 10 MHz), to make the data a bit more manageable.

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.

Alex V Flinsch
Data Analytics – Paper 2

Step 1 – Get the wspr data, and load to SPSS.
I downloaded the data for January 2009 located at http://wsprnet.org/archive/wsprspots-2009-01.csv.zip. This is a comma-delimited dataset which I unzipped and added a header row containing the column names to the top of the file. I added the header row to avoid having to type all of the variable names. After loading the data into SPSS, I added a column called 'date' and transformed the Unix timestamp column into a readable date. I was sort of surprise that SPSS did not have a built-in function or a date format to readily handle this. SPSS did have a handy DATESUM function, which I was able to use by adding the timestamp as seconds from 01/01/1970 (base of the Unix timeline). This worked, and the data was saved. I then did the same for the January 2013 wspr data.

Output:

```
  GET DATA  /TYPE=TXT
  /FILE="/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wsprspo
ts-2009-01.csv"
  /ENCODING='Locale'
  /DELCASE=LINE
  /DELIMITERS=","
  /ARRANGEMENT=DELIMITED
  /FIRSTCASE=2
  /IMPORTCASE=ALL
  /VARIABLES=
  spot_id F7.0
  timestamp F10.0
  rx_station A15
  rx_grid A6
  snr F3.0
  frequency F10.6
  tx_station A15
  tx_grid A6
  tx_power F2.0
  drift F2.0
  distance F5.0
  azimuth F3.0
  band F3.0
  version A10
  code F1.0.
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

SAVE OUTFILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/ws
pr_2009_01.sav'
  /COMPRESSED.
COMPUTE date=DATESUM(Date.Dmy(01,01,1970),timestamp,'seconds').
EXECUTE.
DATASET ACTIVATE DataSet1.

SAVE OUTFILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/ws
pr_2009_01.sav'
  /COMPRESSED.

GET DATA  /TYPE=TXT
  /FILE="/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wsprspo
ts-2013-01.csv"
  /ENCODING='Locale'
  /DELCASE=LINE
  /DELIMITERS=","
  /ARRANGEMENT=DELIMITED
  /FIRSTCASE=2
  /IMPORTCASE=ALL
  /VARIABLES=
  spot_id F9.0
  timestamp F10.0
  rx_station A15
  rx_grid A6
  snr F3.0
  frequency F9.6
  tx_station A15
  tx_grid A6
  tx_power F2.0
  drift F2.0
  distance F5.0
  azimuth F3.0
  band F2.0
  version A10
  code F1.0.
CACHE.
EXECUTE.
DATASET NAME DataSet2 WINDOW=FRONT.
COMPUTE date=DATESUM(DATE.DMY(01,01,1979),timestamp,'seconds').
EXECUTE.
COMPUTE date=DATESUM(DATE.DMY(01,01,1970),timestamp,'seconds').
EXECUTE.

SAVE OUTFILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/ws
pr_2013_01.sav'
  /COMPRESSED.
```

Step 2 – Get solar data and load to SPSS.
The solar data for 2009 was downloaded from http://www.swpc.noaa.gov/ftpdir/warehouse/2009/2009_DSD.txt. This is a fixed column dataset, with several rows of descriptive text above the actual data. I needed to play about with importing the date column which was formatted as 'yyyy mm dd', something which SPSS does not recognize as a date format (another surprise). I was able to import the data as separate year, month, day columns and transform them into a single date field. Once I finished with the 2009 data, I saved a pre defined file format, to allow for easier import on the 2013 data.

Output:

```
NEW FILE.
DATASET NAME DataSet1 WINDOW=FRONT.

GET DATA  /TYPE=TXT
  /FILE="/Users/alex/Desktop/spring 2014/Data Analytics/paper
2/2009_DSD.txt"
  /ENCODING='Locale'
  /FIXCASE=1
  /ARRANGEMENT=FIXED
  /FIRSTCASE=14
  /IMPORTCASE=ALL
  /VARIABLES=
  /1 date 0-9 A10
  flux 10-14 F5.0
  ssn 15-21 F7.0
  V4 22-30 F9.0
  V5 31-37 F7.0
  V6 38-45 F8.0
  V7 46-52 A7
  V8 53-56 F4.0
  V9 57-59 F3.0
  V10 60-62 F3.0
  V11 63-65 F3.0
  V12 66-68 F3.0
  V13 69-71 F3.0
  V14 72-74 F3.0.
CACHE.
EXECUTE.
DATASET NAME DataSet2 WINDOW=FRONT.
DATASET ACTIVATE DataSet1.
DATASET CLOSE DataSet2.

GET DATA  /TYPE=TXT
  /FILE="/Users/alex/Desktop/spring 2014/Data Analytics/paper
2/2009_DSD.txt"
  /ENCODING='Locale'
  /FIXCASE=1
  /ARRANGEMENT=FIXED
  /FIRSTCASE=14
  /IMPORTCASE=ALL
  /VARIABLES=
  /1 year 0-3 A4
  month 4-6 A3
  day 7-9 F3.0
  flux 10-14 F5.0
  sunspot_num 15-21 F7.0
  sunspot_area 22-30 F9.0
  new_regions 31-37 F7.0
  mean_field 38-45 F8.0
  xray_flux 46-52 A7
  xc_flare 53-56 F4.0
  xm_flare 57-59 F3.0
  xx_flare 60-62 F3.0
  s_flare 63-65 F3.0
  o1_flare 66-68 F3.0
  o2_flare 69-71 F3.0
  o3_flare 72-74 F3.0.
CACHE.
EXECUTE.
DATASET NAME DataSet3 WINDOW=FRONT.
DATASET CLOSE DataSet1.
COMPUTE date=DATE.MDY(month,day,year).
EXECUTE.

SAVE OUTFILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper
2/solar_2009.sav'
  /COMPRESSED.

GET DATA  /TYPE=TXT
  /FILE="/Users/alex/Desktop/spring 2014/Data Analytics/paper
2/2013_DSD.txt"
  /ENCODING='Locale'
  /FIXCASE=1
  /ARRANGEMENT=FIXED
  /FIRSTCASE=14
  /IMPORTCASE=ALL
  /VARIABLES=
  /1 year 0-3 A4
  month 4-6 A3
  day 7-9 F3.0
  flux 10-14 F5.0
  sunspot_num 15-21 F7.0
  sunspot_area 22-30 F9.0
  new_regions 31-37 F7.0
  mean_field 38-45 F8.0
  xray_flux 46-52 A7
  xc_flare 53-56 F4.0
  xm_flare 57-59 F3.0
  xx_flare 60-62 F3.0
  s_flare 63-65 F3.0
  o1_flare 66-68 F3.0
  o2_flare 69-71 F3.0
  o3_flare 72-74 F3.0.
CACHE.
EXECUTE.
DATASET NAME DataSet4 WINDOW=FRONT.
COMPUTE date=DATE.MDY(month,day,year).
EXECUTE.

SAVE OUTFILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper
2/solar_2013.sav'
  /COMPRESSED.
```

Step 3 – Merging datasets.

This step involves adding the solar data to the wspr data for 2009. The solar data is already in date sequence, but I was not sure if the wspr data was. The wspr data was sorted in ascending sequence by date to match the solar data, and then the datasets were merged and resaved under a different name. This process was repeated with the 2013 data.

Output:

```
GET
  FILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr_200
9_01.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
SORT CASES BY date(A).
MATCH FILES /FILE=*
  /FILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/solar_2
009.sav'
  /RENAME (date = d0)
  /DROP= d0.
EXECUTE.

SAVE OUTFILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/ws
pr+solar_2009_01.sav'
  /COMPRESSED.
GET
  FILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr_201
3_01.sav'.
DATASET NAME DataSet2 WINDOW=FRONT.
MATCH FILES /FILE=*
  /FILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/solar_2
013.sav'
  /RENAME (date = d0)
  /DROP= d0.
EXECUTE.
```

Alex V Flinsch

Data Analytics – Paper 2

Step 4 – Initial data analysis

Limiting the data to a single band (frequency=10 MHz, wavelength=30m), some descriptive information was gathered based on the January 2009 data.

Output:

```
USE ALL.
COMPUTE filter_$=(band=10).
VARIABLE LABELS filter_$ 'band=10 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
DESCRIPTIVES VARIABLES=distance
  /STATISTICS=MEAN STDDEV VARIANCE RANGE MIN MAX.
```

Descriptives

Notes

Output Created		08-MAR-2014 08:30:31
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2009_01.sav
	Active Dataset	DataSet1
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	340956
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	All non-missing data are used.
Syntax		DESCRIPTIVES VARIABLES=distance /STATISTICS=MEAN STDDEV VARIANCE RANGE MIN MAX.
Resources	Processor Time	00:00:00.41
	Elapsed Time	00:00:01.00

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
distance	340956	19929	0	19929	2962.96	3564.457	12705351.479
Valid N (listwise)	340956						

Step 5 – Time for some graphs January 2009

Still limiting the data to a single band (frequency=10 MHz, wavelength=30m), a timeline of mean distance was plotted and using only the solar data, a timeline of sunspot count was also generated.

Output:

```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=date
MEAN(distance)[name="MEAN_distance"] MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: date=col(source(s), name("date"))
  DATA: MEAN_distance=col(source(s), name("MEAN_distance"))
  GUIDE: axis(dim(1), label("date"))
  GUIDE: axis(dim(2), label("Mean distance"))
  GUIDE: text.title(label("Mean Distance on 10MHz January 2009"))
  ELEMENT: point(position(date*MEAN_distance))
END GPL.
```

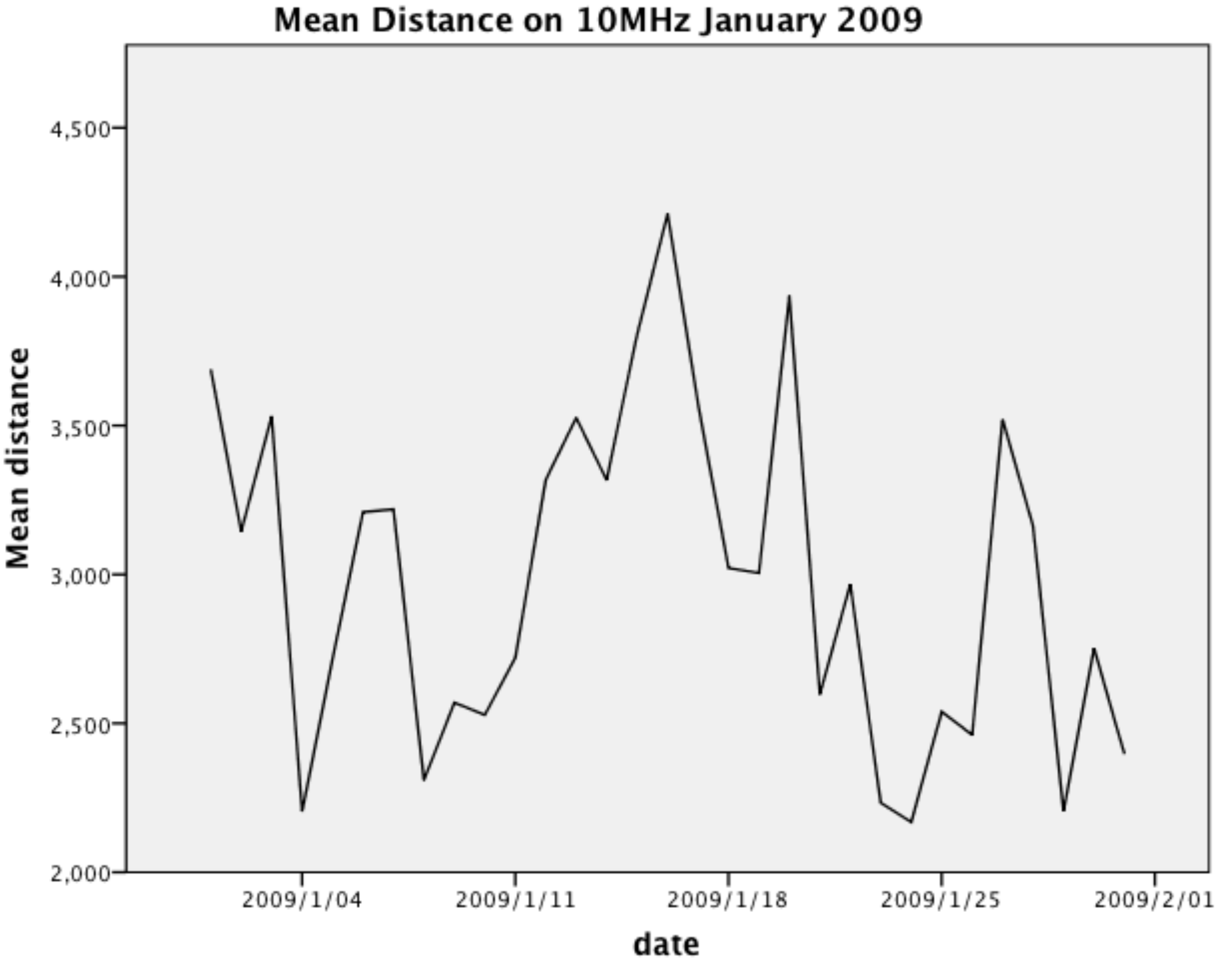
GGraph

Notes

Output Created		05-MAR-2014 19:06:27
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2009_01.sav
	Active Dataset	DataSet1
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	340956

Syntax		GGRAPH
		  /GRAPHDATASET NAME="graphdataset" VARIABLES=date MEAN(distance)[name="MEAN_distance"] MISSING=LISTWISE REPORTMISSING=NO
		  /GRAPHSPEC SOURCE=INLINE.
		BEGIN GPL
		  SOURCE: s=userSource(id("graphdataset"))
		  DATA: date=col(source(s), name("date"))
		  DATA: MEAN_distance=col(source(s), name("MEAN_distance"))
		  GUIDE: axis(dim(1), label("date"))
		  GUIDE: axis(dim(2), label("Mean distance"))
		  GUIDE: text.title(label("Mean Distance on 10MHz January 2009"))
		  ELEMENT: point(position(date*MEAN_distance))
		END GPL.

Resources	Processor Time	00:00:00.76
	Elapsed Time	00:00:01.00


Mean Distance on 10MHz January 2009
Mean distance
4,500
4,000
3,500
3,000
2,500
2,000
2009/1/04
2009/1/11
2009/1/18
2009/1/25
2009/2/01
date

Alex V Flinsch
Data Analytics – Paper 2

```
DESCRIPTIVES VARIABLES=distance
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

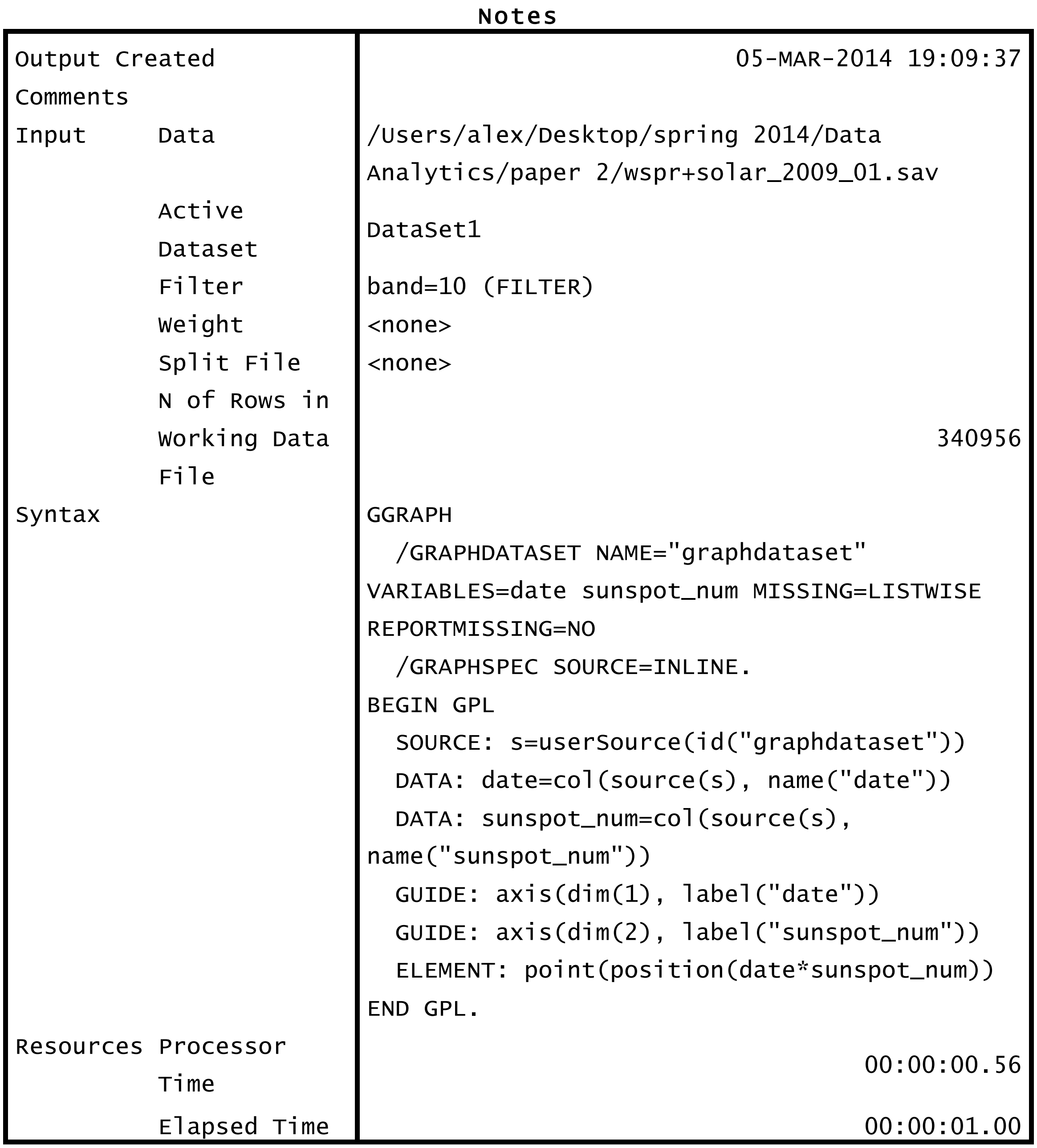
Notes

Output Created		05-MAR-2014 19:08:03
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2009_01.sav
	Active Dataset	DataSet1
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	340956
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	All non-missing data are used.
Syntax		DESCRIPTIVES VARIABLES=distance /STATISTICS=MEAN STDDEV MIN MAX.
Resources	Processor Time	00:00:00.32
	Elapsed Time	00:00:00.00

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
distance	340956	0	19929	2962.96	3564.457
Valid N (listwise)	340956				

Alex V Flinsch
Data Analytics – Paper 2



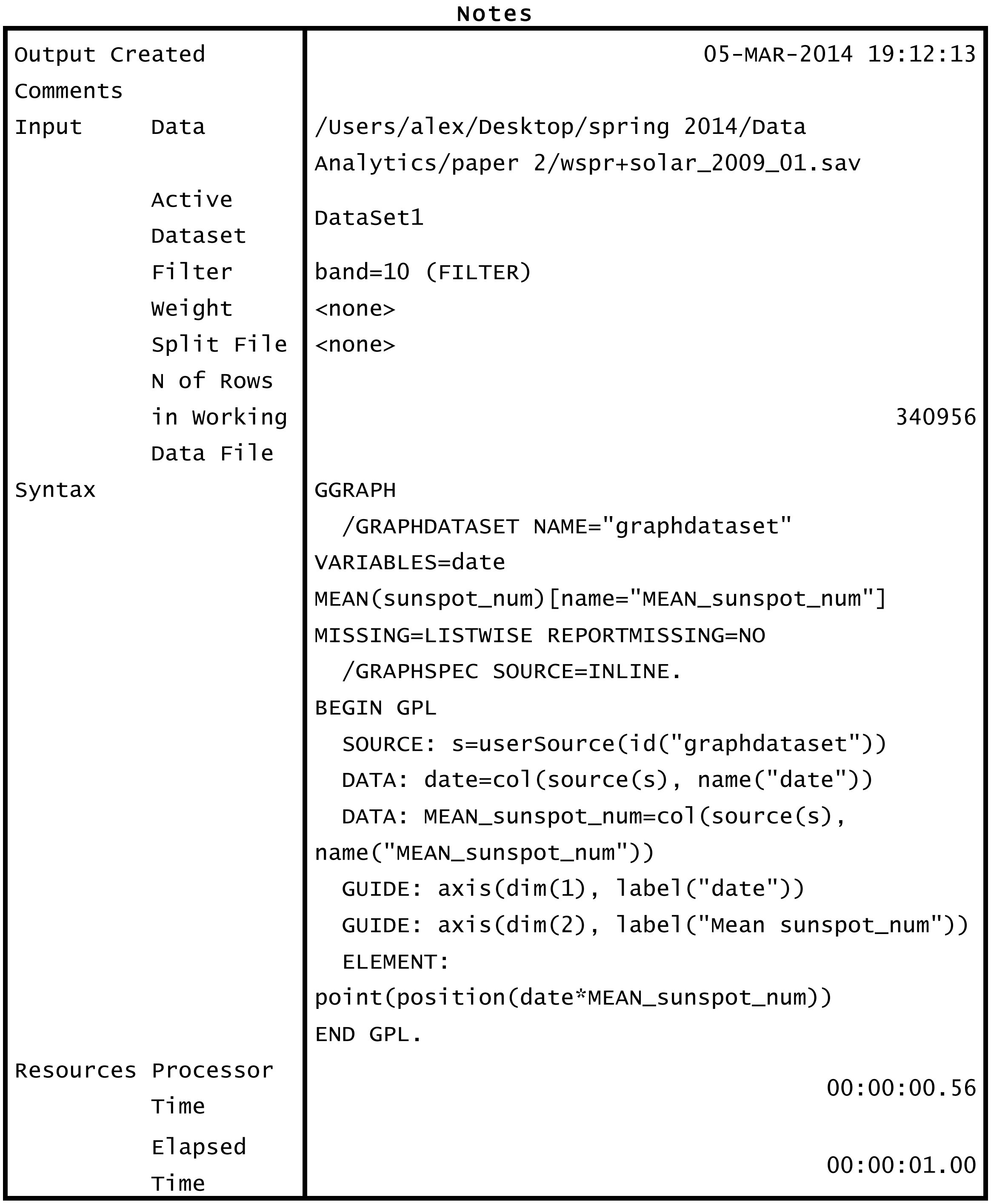
Notes

Output Created		05-MAR-2014 19:09:37
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2009_01.sav
	Active Dataset	DataSet1
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	340956

Syntax		GGRAPH
		  /GRAPHDATASET NAME="graphdataset" VARIABLES=date sunspot_num MISSING=LISTWISE REPORTMISSING=NO
		  /GRAPHSPEC SOURCE=INLINE.
		BEGIN GPL
		  SOURCE: s=userSource(id("graphdataset"))
		  DATA: date=col(source(s), name("date"))
		  DATA: sunspot_num=col(source(s), name("sunspot_num"))
		  GUIDE: axis(dim(1), label("date"))
		  GUIDE: axis(dim(2), label("sunspot_num"))
		  ELEMENT: point(position(date*sunspot_num))
		END GPL.

Resources	Processor Time	00:00:00.56
	Elapsed Time	00:00:01.00

Alex V Flinsch
Data Analytics – Paper 2



Notes

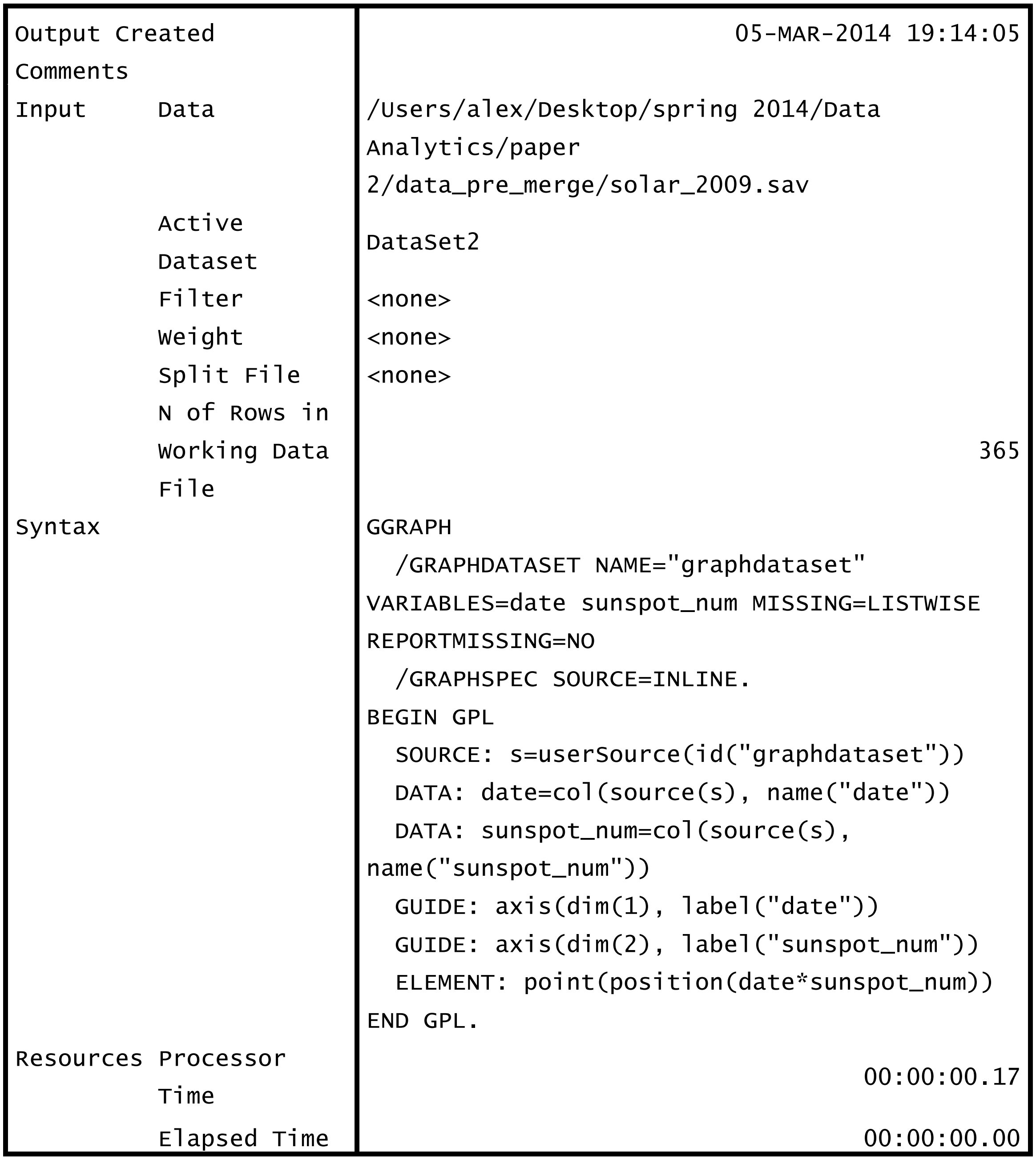
Output Created		05-MAR-2014 19:12:13
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2009_01.sav
	Active Dataset	DataSet1
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	340956

Syntax		GGRAPH
		  /GRAPHDATASET NAME="graphdataset" VARIABLES=date MEAN(sunspot_num)[name="MEAN_sunspot_num"] MISSING=LISTWISE REPORTMISSING=NO
		  /GRAPHSPEC SOURCE=INLINE.
		BEGIN GPL
		  SOURCE: s=userSource(id("graphdataset"))
		  DATA: date=col(source(s), name("date"))
		  DATA: MEAN_sunspot_num=col(source(s), name("MEAN_sunspot_num"))
		  GUIDE: axis(dim(1), label("date"))
		  GUIDE: axis(dim(2), label("Mean sunspot_num"))
		  ELEMENT: point(position(date*MEAN_sunspot_num))
		END GPL.

Resources	Processor Time	00:00:00.56
	Elapsed Time	00:00:01.00

Alex V Flinsch
Data Analytics – Paper 2

Notes



Output Created		05-MAR-2014 19:14:05
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/data_pre_merge/solar_2009.sav
	Active Dataset	DataSet2
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	365

Syntax		GGRAPH
		  /GRAPHDATASET NAME="graphdataset" VARIABLES=date sunspot_num MISSING=LISTWISE REPORTMISSING=NO
		  /GRAPHSPEC SOURCE=INLINE.
		BEGIN GPL
		  SOURCE: s=userSource(id("graphdataset"))
		  DATA: date=col(source(s), name("date"))
		  DATA: sunspot_num=col(source(s), name("sunspot_num"))
		  GUIDE: axis(dim(1), label("date"))
		  GUIDE: axis(dim(2), label("sunspot_num"))
		  ELEMENT: point(position(date*sunspot_num))
		END GPL.

Resources	Processor Time	00:00:00.17
	Elapsed Time	00:00:00.00

Notes

Output Created		05-MAR-2014 19:16:55
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/data_pre_merge/solar_2009.sav
	Active Dataset	DataSet2
	Filter	month=1 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	31

Syntax		GGRAPH
		  /GRAPHDATASET NAME="graphdataset" VARIABLES=date sunspot_num MISSING=LISTWISE REPORTMISSING=NO
		  /GRAPHSPEC SOURCE=INLINE.
		BEGIN GPL
		  SOURCE: s=userSource(id("graphdataset"))
		  DATA: date=col(source(s), name("date"))
		  DATA: sunspot_num=col(source(s), name("sunspot_num"))
		  GUIDE: axis(dim(1), label("date"))
		  GUIDE: axis(dim(2), label("sunspot_num"))
		  TRANS: date_sunspot_num=eval("date - sunspot_num")
		  TRANS: date_sunspot_num=eval("date - sunspot_num")
		  ELEMENT: point(position(date*sunspot_num), color.exterior(date_sunspot_num))
		  ELEMENT: point(position(date*sunspot_num), color.exterior(date_sunspot_num))
		END GPL.

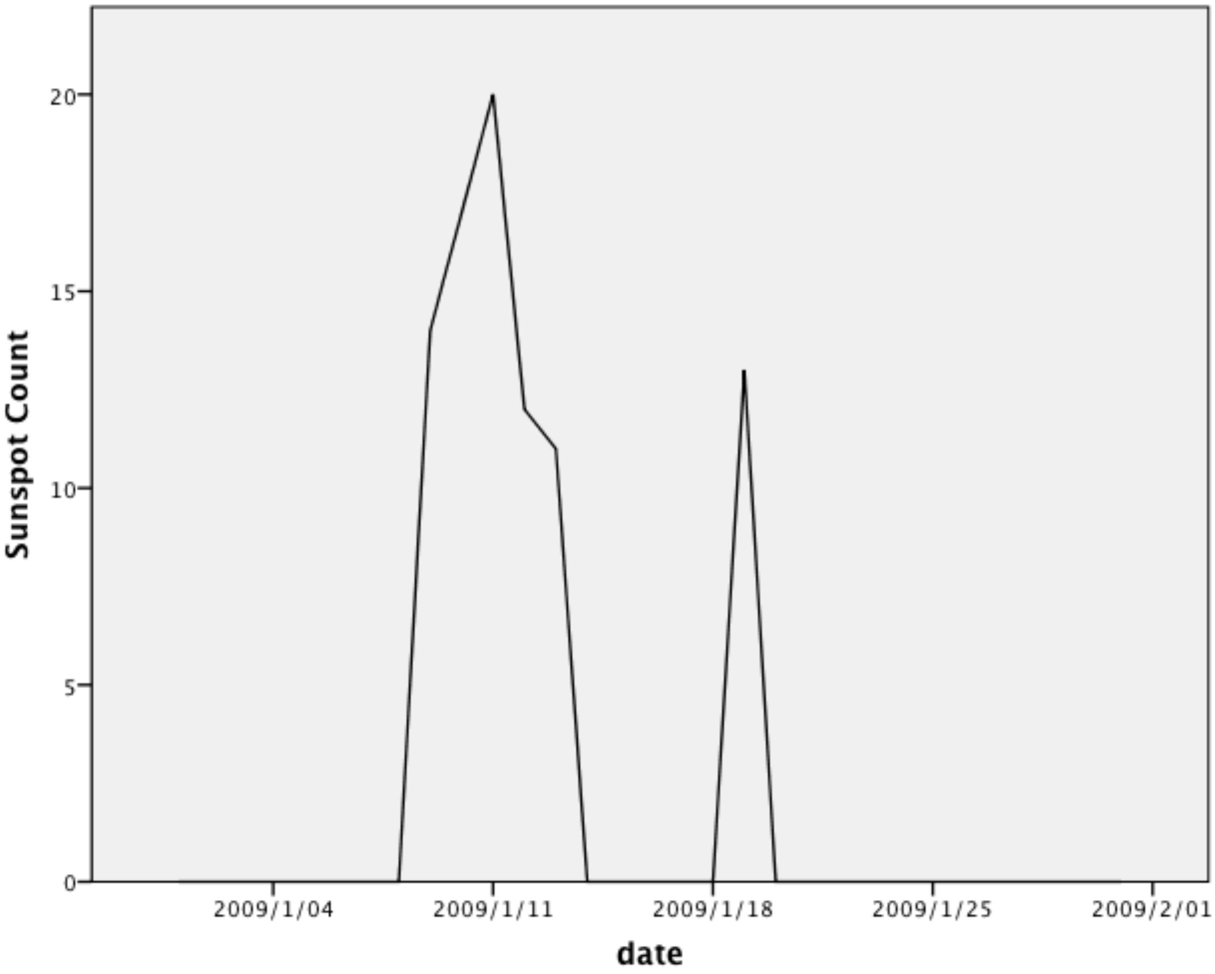
Resources	Processor Time	00:00:00.18
	Elapsed Time	00:00:00.00

```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=date sunspot_num
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: date=col(source(s), name("date"))
  DATA: sunspot_num=col(source(s), name("sunspot_num"))
  GUIDE: axis(dim(1), label("date"))
  GUIDE: axis(dim(2), label("sunspot_num"))
  ELEMENT: point(position(date*sunspot_num))
END GPL.
```

GGraph

Notes

Output Created		05-MAR-2014 19:20:02
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/data_pre_merge/solar_2009.sav
	Active Dataset	DataSet2
	Filter	month=1 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	31
Syntax		GGRAPH /GRAPHDATASET NAME="graphdataset" VARIABLES=date sunspot_num MISSING=LISTWISE REPORTMISSING=NO /GRAPHSPEC SOURCE=INLINE. BEGIN GPL SOURCE: s=userSource(id("graphdataset")) DATA: date=col(source(s), name("date")) DATA: sunspot_num=col(source(s), name("sunspot_num")) GUIDE: axis(dim(1), label("date")) GUIDE: axis(dim(2), label("sunspot_num")) ELEMENT: point(position(date*sunspot_num)) END GPL.
Resources	Processor Time	00:00:00.13
	Elapsed Time	00:00:00.00


Sunspot Count
20
15
10
5
0
2009/1/04
2009/1/11
2009/1/18
2009/1/25
2009/2/01
date

Step 6 – Looking at the data graphically – January 2009

One of the first things I noticed was that there were two distinct spikes in the number of sunspots on a daily basis during the month. This is what I had expected before plotting the data, as a quick 'eyeball' of the data showed that there were several periods in the month with a zero sunspot count. When comparing the two plots (sunspots vs date and mean distance vs date) there were distinct spikes in the mean distance a few days after an increase in the sunspot count. Again, this is what I had expected based on the travel time for ionized particles emitted from the sunspots to reach the Earth and interact with the ionosphere. There was a peak distance at the beginning of the month and I suspected that there might have been an increase in sunspots at the end of December 2008; an examination of the 2008 data shows that NOT to be the case.

Step 7 – Looking at a different set of data – January 2013

Using the same selection of data, but from January 2013, the same plots were generated.

Output:

```
GET
  FILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper
2/wspr+solar_2013_01.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
USE ALL.
COMPUTE filter_$=(band=10).
VARIABLE LABELS filter_$ 'band=10 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=date distance
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: date=col(source(s), name("date"))
  DATA: distance=col(source(s), name("distance"))
  GUIDE: axis(dim(1), label("date"))
  GUIDE: axis(dim(2), label("distance"))
  ELEMENT: point(position(date*distance))
END GPL.
```

GGraph

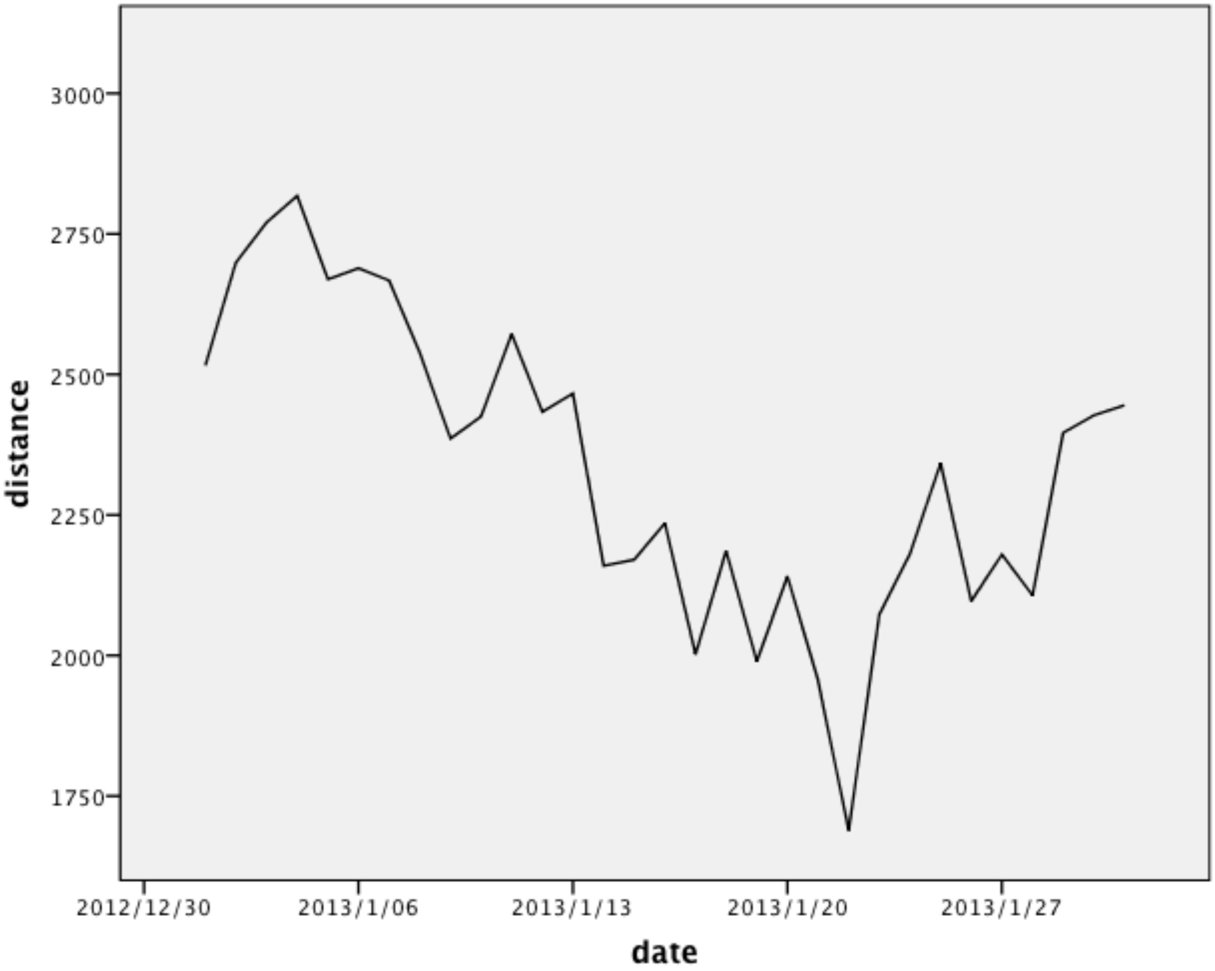
Notes

Output Created		05-MAR-2014 21:30:33
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2013_01.sav
	Active Dataset	DataSet1
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	813252

Syntax		GGRAPH
		  /GRAPHDATASET NAME="graphdataset" VARIABLES=date distance MISSING=LISTWISE REPORTMISSING=NO
		  /GRAPHSPEC SOURCE=INLINE.
		BEGIN GPL
		  SOURCE: s=userSource(id("graphdataset"))
		  DATA: date=col(source(s), name("date"))
		  DATA: distance=col(source(s), name("distance"))
		  GUIDE: axis(dim(1), label("date"))
		  GUIDE: axis(dim(2), label("distance"))
		  ELEMENT: point(position(date*distance))
		END GPL.

Resources	Processor Time	00:00:05.67
	Elapsed Time	00:00:06.00

```
[DataSet1] /Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2013_01.sav
```


3000
2750
2500
2250
2000
1750
distance
2012/12/30
2013/1/06
2013/1/13
2013/1/20
2013/1/27
date

Alex V Flinsch
Data Analytics – Paper 2

```
DESCRIPTIVES VARIABLES=distance
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

Notes

Output Created		05-MAR-2014 21:35:57
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2013_01.sav
	Active Dataset	DataSet1
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	813252
Missing Value Handling	Definition of Missing	User defined missing values are treated as missing.
	Cases Used	All non-missing data are used.
Syntax		DESCRIPTIVES VARIABLES=distance /STATISTICS=MEAN STDDEV MIN MAX.
Resources	Processor Time	00:00:02.12
	Elapsed Time	00:00:02.00

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
distance	813252	0	19468	2356.60	2721.597
Valid N (listwise)	813252				

```
GET
  FILE='/Users/alex/Desktop/spring 2014/Data Analytics/paper
2/data_pre_merge/solar_2013.sav'.
DATASET NAME DataSet2 WINDOW=FRONT.
USE ALL.
COMPUTE filter_$=(month=1).
VARIABLE LABELS filter_$ 'month=1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=date sunspot_num
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: date=col(source(s), name("date"))
  DATA: sunspot_num=col(source(s), name("sunspot_num"))
  GUIDE: axis(dim(1), label("date"))
  GUIDE: axis(dim(2), label("sunspot_num"))
  ELEMENT: point(position(date*sunspot_num))
END GPL.
```

Alex V Flinsch
Data Analytics – Paper 2

GGraph

Notes

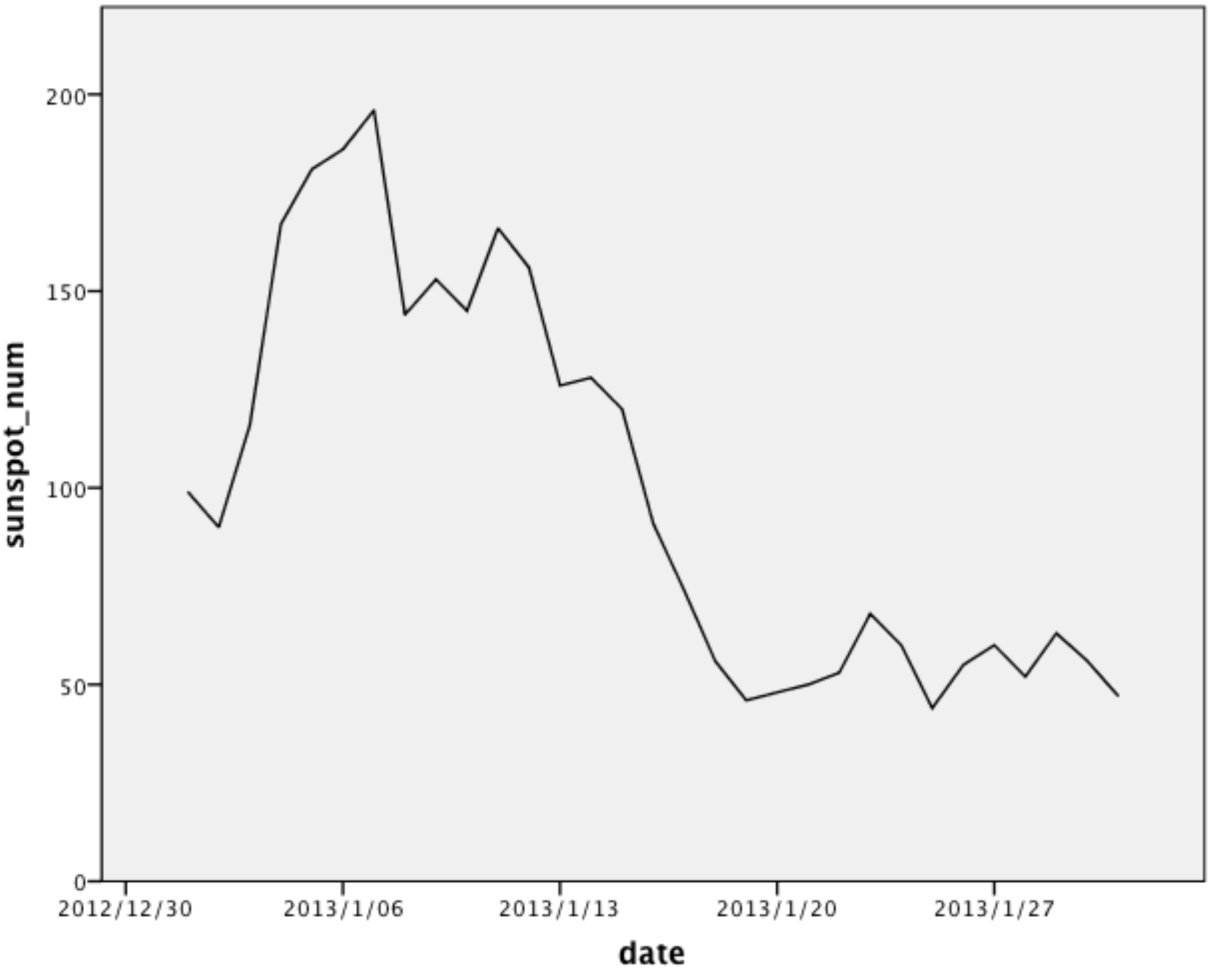
Output Created		05-MAR-2014 21:37:39
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/data_pre_merge/solar_2013.sav
	Active Dataset	DataSet2
	Filter	month=1 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	31

Syntax		GGRAPH
		  /GRAPHDATASET NAME="graphdataset" VARIABLES=date sunspot_num MISSING=LISTWISE REPORTMISSING=NO
		  /GRAPHSPEC SOURCE=INLINE.
		BEGIN GPL
		  SOURCE: s=userSource(id("graphdataset"))
		  DATA: date=col(source(s), name("date"))
		  DATA: sunspot_num=col(source(s), name("sunspot_num"))
		  GUIDE: axis(dim(1), label("date"))
		  GUIDE: axis(dim(2), label("sunspot_num"))
		  ELEMENT: point(position(date*sunspot_num))
		END GPL.

Resources	Processor Time	00:00:00.15
	Elapsed Time	00:00:00.00

```
[DataSet2] /Users/alex/Desktop/spring 2014/Data Analytics/paper
2/data_pre_merge/solar_2013.sav
```

Alex V Flinsch
Data Analytics – Paper 2



Step 8 – Looking at the data graphically – January 2013

Immediately I noticed was that the overall shapes of the two graphs were fairly similar, the mean distance a signal was reported to travel appears to be very closely linked to the number of sunspots.

Step 9 – Linear Regression – January 2009

Using distance as the dependent, and sunspot count as zpredictor variables
Output:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT distance
  /METHOD=ENTER sunspot_num
  /SCATTERPLOT=(distance ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

Alex V Flinsch
Data Analytics – Paper 2

Regression

Notes

Output Created		08-MAR-2014 10:15:16
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2009_01.sav
	Active Dataset	DataSet1
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	340956
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
Syntax		REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT distance /METHOD=ENTER sunspot_num /SCATTERPLOT=(distance ,*ZPRED) /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
Resources	Processor Time	00:00:01.68
	Elapsed Time	00:00:01.00
	Memory Required	3760 bytes
	Additional Memory Required for Residual Plots	896 bytes

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	sunspot_num[b]	.	Enter

a. Dependent Variable: distance

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.058[a]	.003	-.001	5713.011

a. Predictors: (Constant), sunspot_num

b. Dependent Variable: distance

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	26719741.809	1	26719741.809	.819	.366[b]
	Residual	7865878271.154	241	32638499.050		
	Total	7892598012.963	242			

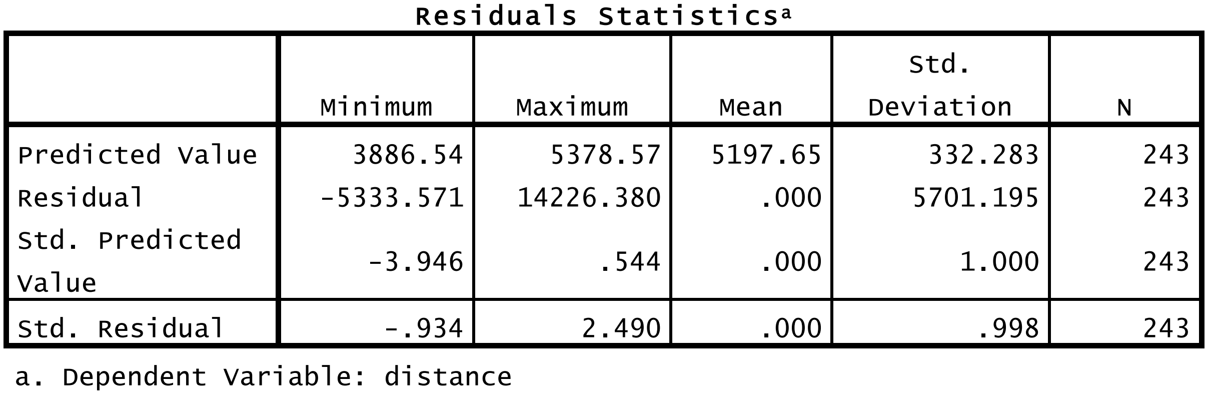
a. Dependent Variable: distance

b. Predictors: (Constant), sunspot_num

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5378.571	417.488		12.883	.000
	sunspot_num	-34.698	38.349	-.058	-.905	.366

a. Dependent Variable: distance

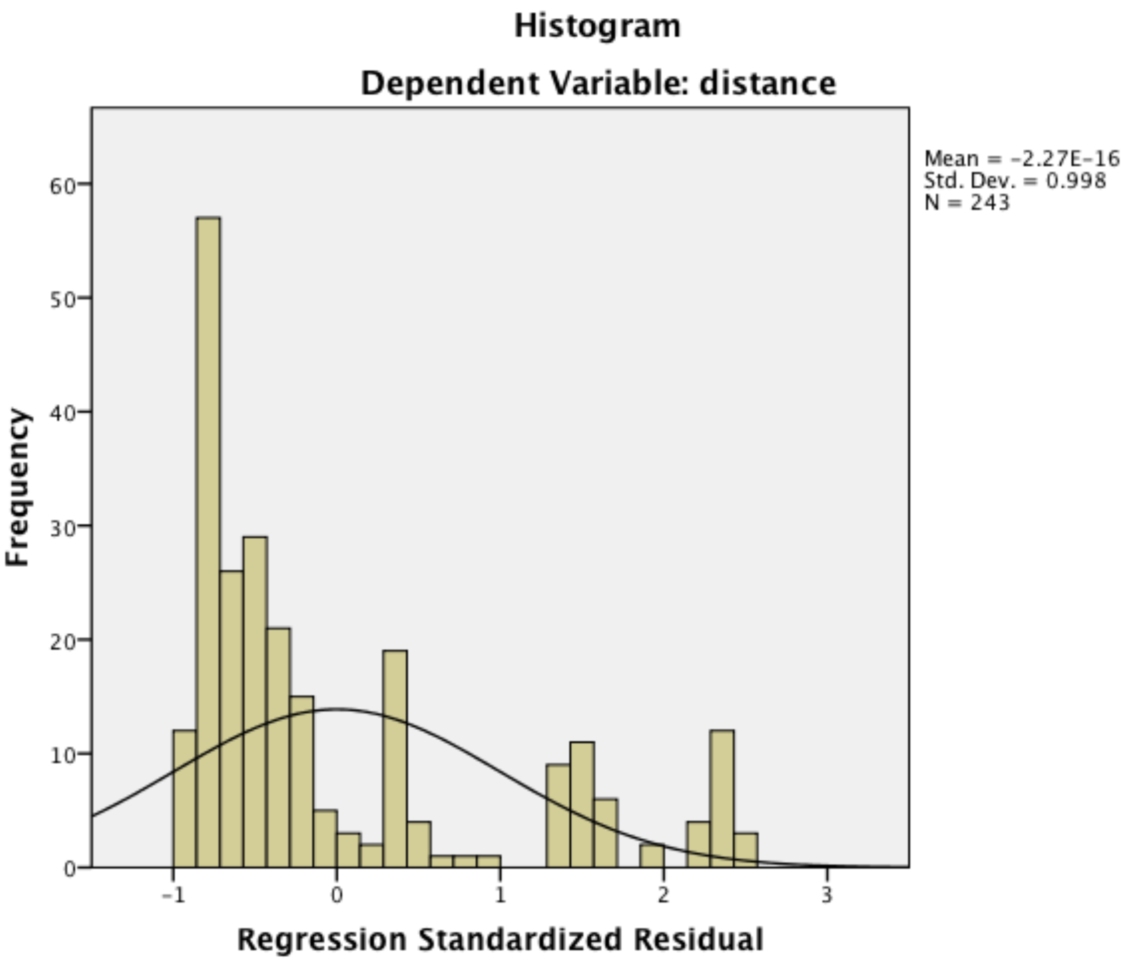


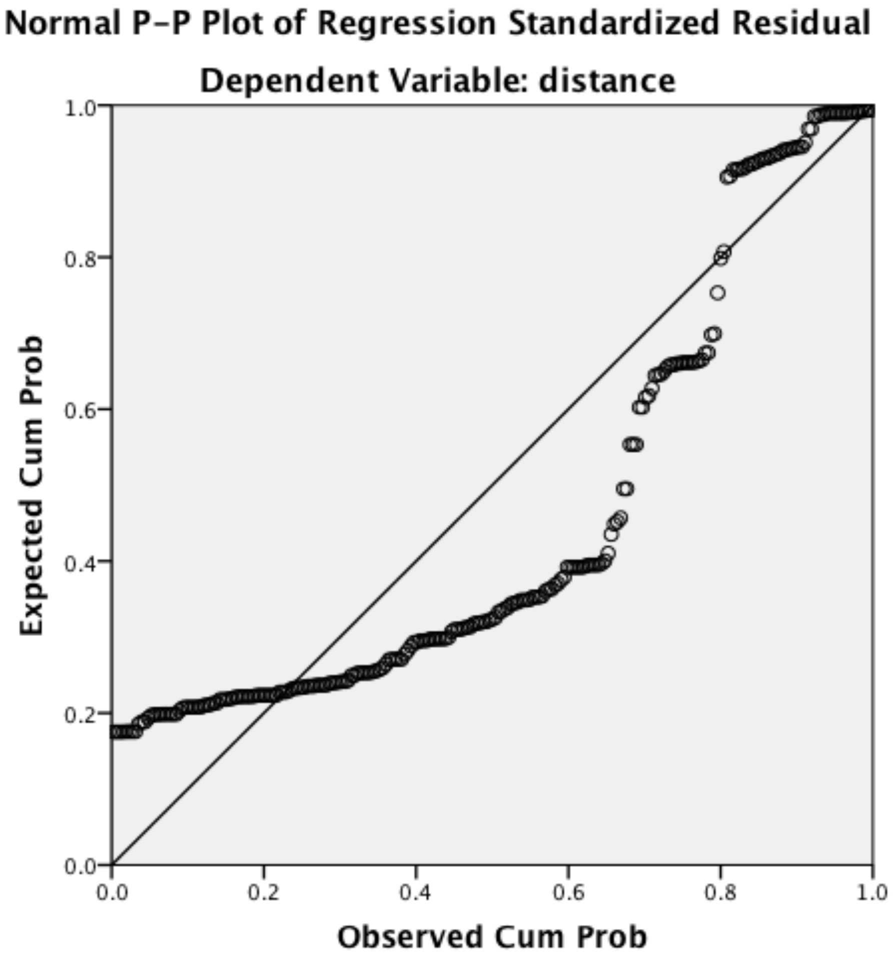
Residuals Statistics[a]

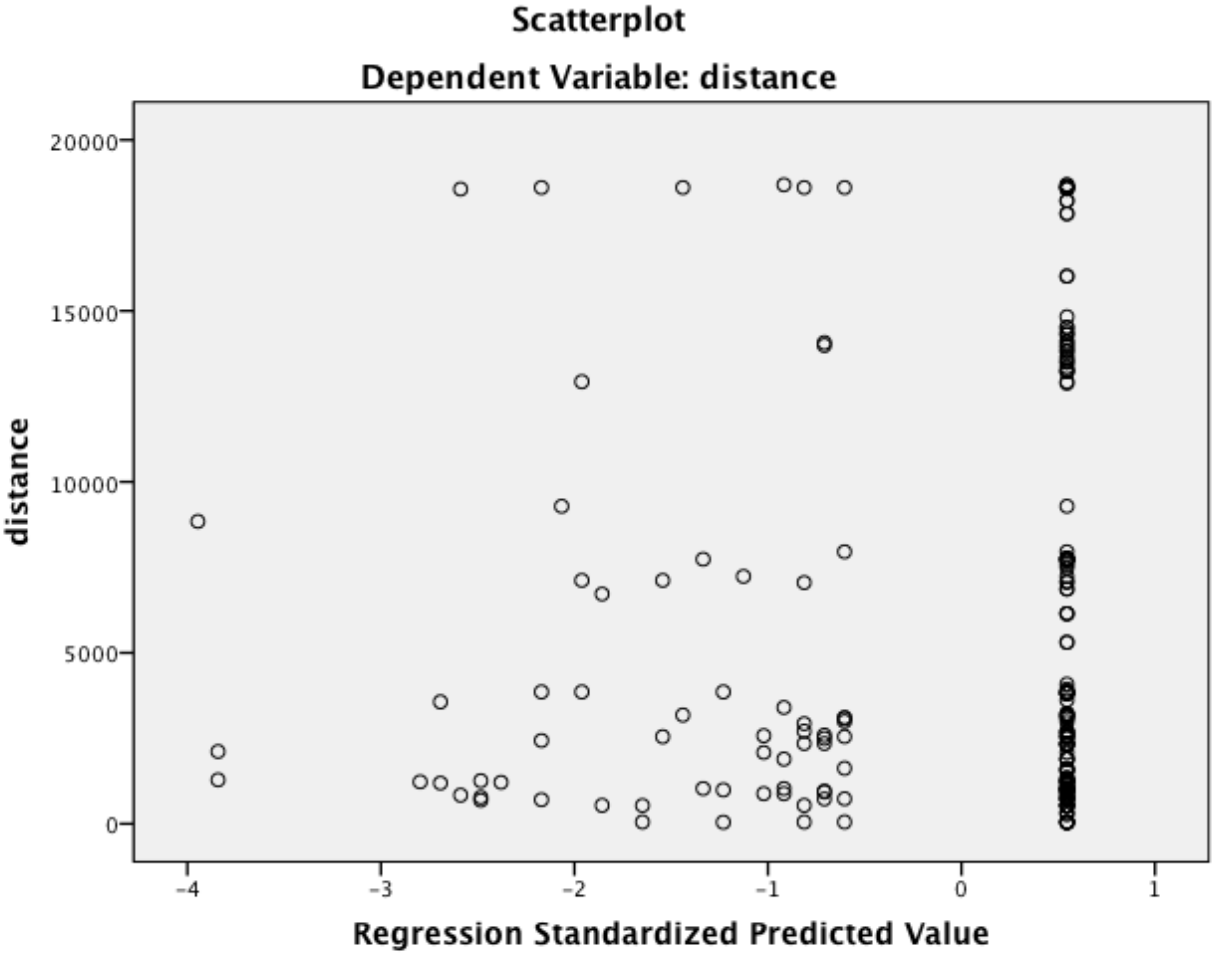
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3886.54	5378.57	5197.65	332.283	243
Residual	-5333.571	14226.380	.000	5701.195	243
Std. Predicted Value	-3.946	.544	.000	1.000	243
Std. Residual	-.934	2.490	.000	.998	243

a. Dependent Variable: distance

Charts


Histogram
Dependent Variable: distance
Mean = -2.27E-16
Std. Dev. = 0.998
N = 243
Frequency
0
10
20
30
40
50
60
-1
0
1
2
3
Regression Standardized Residual


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: distance
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0


Scatterplot
Dependent Variable: distance
distance
20000
15000
10000
5000
0
-4
-3
-2
-1
0
1
Regression Standardized Predicted Value

Step 10 – Linear Regression – January 2013

Using distance as the dependent, and sunspot count as zpredictor variables

Output:

```
USE ALL.
COMPUTE filter_$=(band=10).
VARIABLE LABELS filter_$ 'band=10 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT distance
  /METHOD=ENTER sunspot_num
  /SCATTERPLOT=(distance ,*ZPRED)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

Regression

Notes

Output Created		08-MAR-2014 10:17:50
Comments		
Input	Data	/Users/alex/Desktop/spring 2014/Data Analytics/paper 2/wspr+solar_2013_01.sav
	Active Dataset	DataSet2
	Filter	band=10 (FILTER)
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	813252
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
Syntax		REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT distance /METHOD=ENTER sunspot_num /SCATTERPLOT=(distance ,*ZPRED) /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
Resources	Processor Time	00:00:06.51
	Elapsed Time	00:00:06.00
	Memory Required	3760 bytes
	Additional Memory Required for Residual Plots	896 bytes

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	sunspot_num[b]	.	Enter

a. Dependent Variable: distance

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.100[a]	.010	-.002	1726.382

a. Predictors: (Constant), sunspot_num

b. Dependent Variable: distance

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2582952.871	1	2582952.871	.867	.355[b]
	Residual	253333586.646	85	2980395.137		
	Total	255916539.517	86			

a. Dependent Variable: distance

b. Predictors: (Constant), sunspot_num

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1916.068	582.821		3.288	.001
	sunspot_num	6.101	6.554	.100	.931	.355

a. Dependent Variable: distance

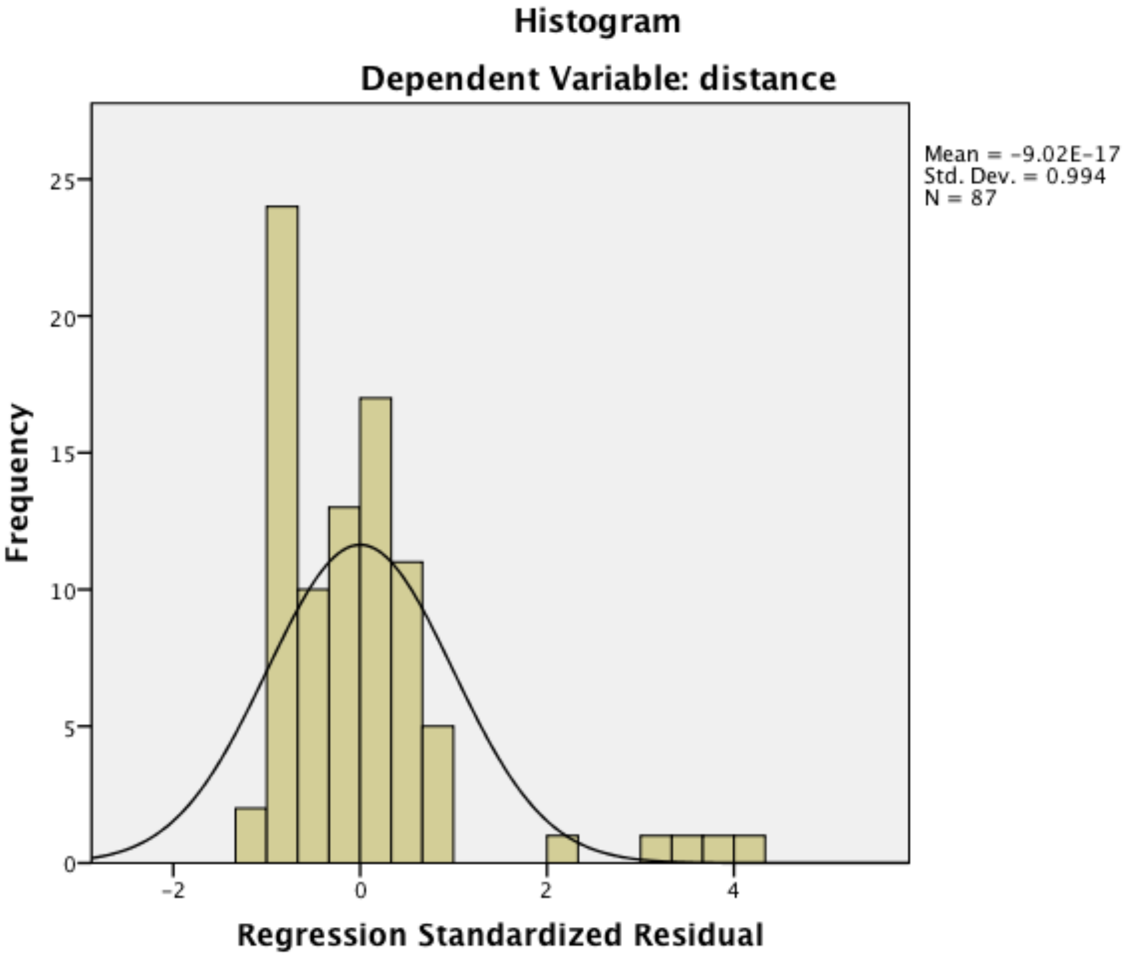
Residuals Statistics[a]

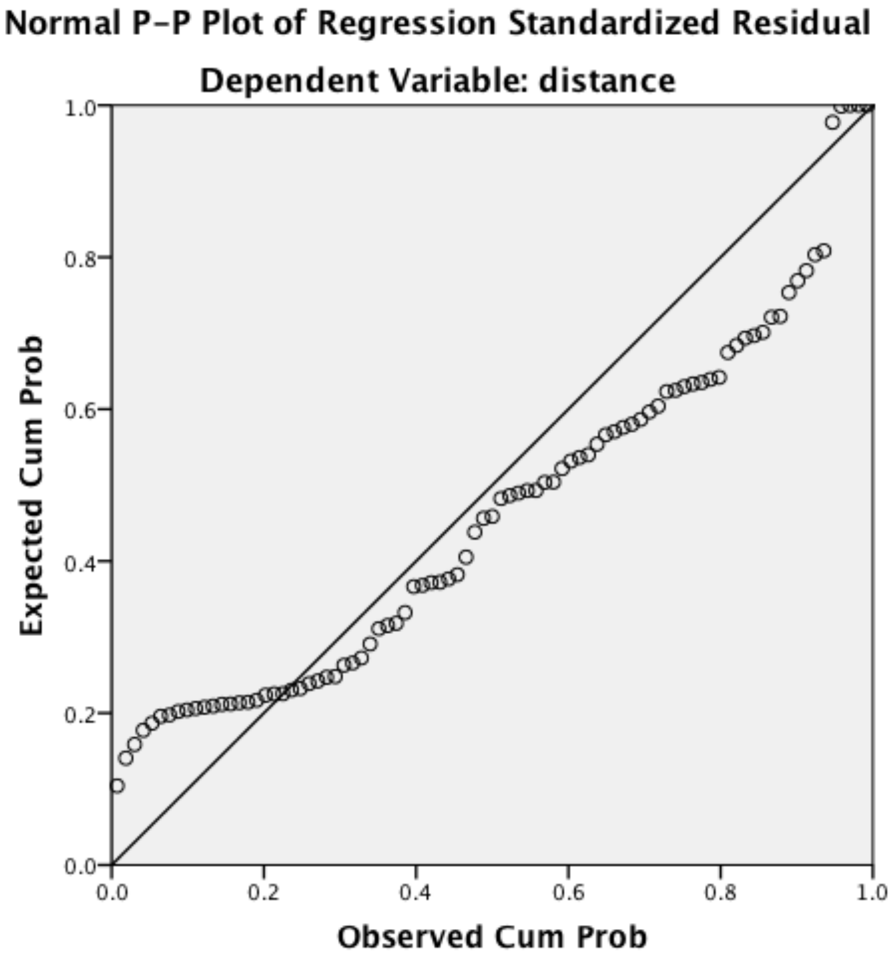
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2001.49	2794.67	2430.55	173.304	87
Residual	-2171.617	7305.325	.000	1716.316	87
Std. Predicted Value	-2.476	2.101	.000	1.000	87
Std. Residual	-1.258	4.232	.000	.994	87

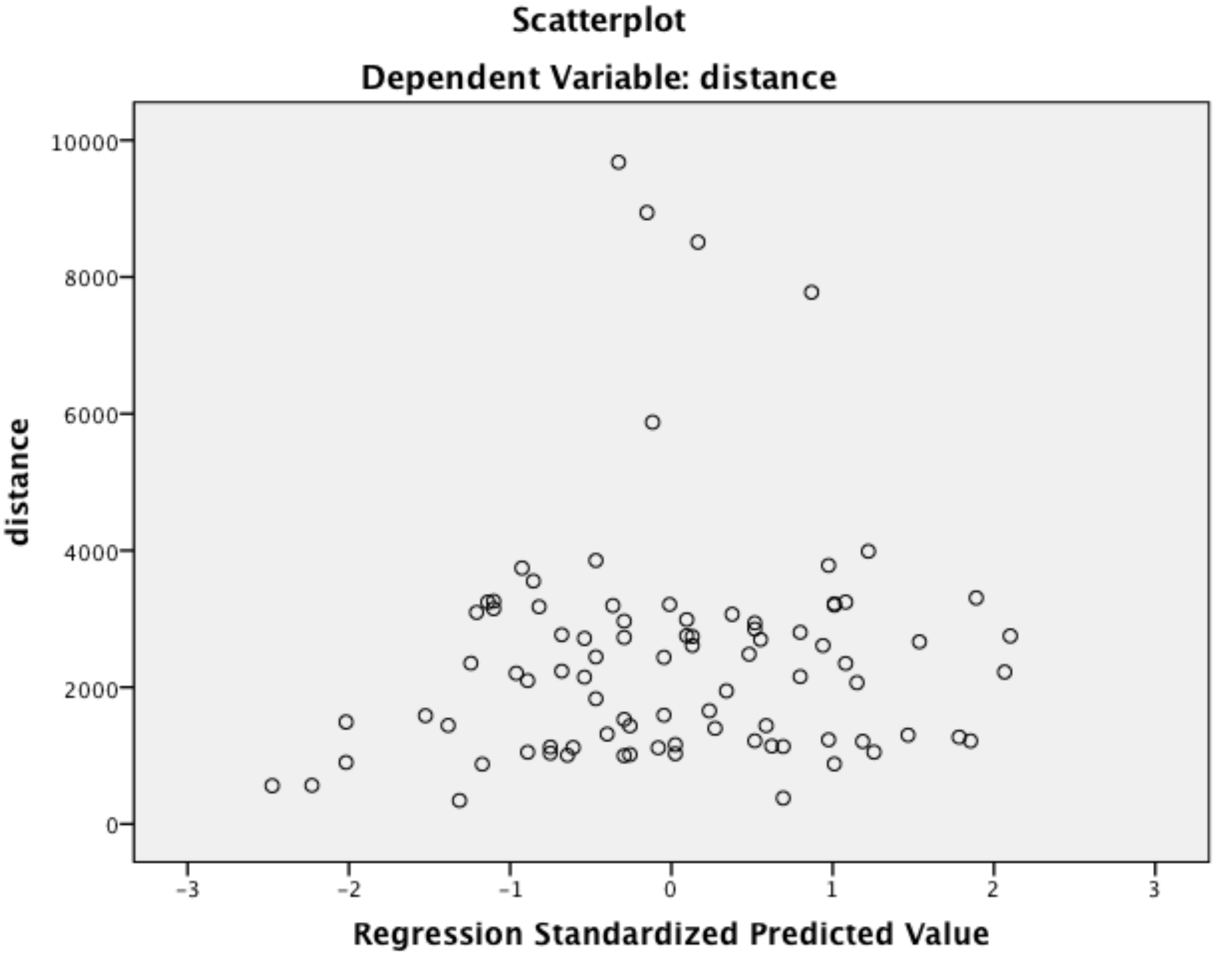
a. Dependent Variable: distance

Alex V Flinsch
Data Analytics – Paper 2

Charts




Normal P-P Plot of Regression Standardized Residual
Dependent Variable: distance
Expected Cum Prob
Observed Cum Prob
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Scatterplot
Dependent Variable: distance
distance
10000
8000
6000
4000
2000
0
-3
-2
-1
0
1
2
3
Regression Standardized Predicted Value

Step 11 – A closer look at the data

Descriptive Statistics - January 2009

	N	Minimum	Maximum	Mean	Std. Deviation
distance	340956	0	19929	2962.96	3564.457
Valid N (listwise)	340956				

Descriptive Statistics - January 2013

	N	Minimum	Maximum	Mean	Std. Deviation
distance	813252	0	19468	2356.60	2721.597
Valid N (listwise)	813252				

The maximum distances for both months appear to be about correct, the values are close to the antipodal distance on Earth (about 20,000 km) – you just can't get any farther apart than that. The minimum distance did not appear to be correct, a signal going nowhere just can't happen. Sorting the data by distance, I saw that there were quite a few cases where the distance was 0. Taking one of the cases (sets of reports between stations W3YTS and W3PRB), I saw that both are located in grid square FM18kx. After looking up the addresses of the stations in the FCC ULS database, I checked Google Maps to see how far apart they were – the driving distance turned out to be less than 4 miles. Since the grid squares used were the same, and no two points in any 6 position locator can be more than 12km apart, the 4 miles, while not zero, it is within the error range or the distance calculation.

In my initial paper, I stated that I should probably eliminate any case where the distance was less than 150 km. I did not do that for this initial data analysis attempt. Since I already had the data for both January 2009, and January 2013 in a local MySQL database, I did a simple query to see how many cases there were

```
select count(*) from spots where
distance <= 150
and band = 10
and date between '2009/01/01' and '2009/01/31'; returned 19312
```

and

```
select count(*) from spots where
distance <= 150
and band = 10
and date between '2009/01/01' and '2009/01/31';  returned 15985
```

Rerunning the descriptive statistics for both sets of data results in:

Descriptive Statistics - January 2009

	N	Range	Minimum	Maximum	Mean	Std. Deviation
distance	321606	19778	151	19929	3139.03	3594.924
Valid N (listwise)	321606					

Descriptive Statistics - January 2013

	N	Range	Minimum	Maximum	Mean	Std. Deviation
distance	797220	19317	151	19468	2402.99	2728.889
Valid N (listwise)	797220					

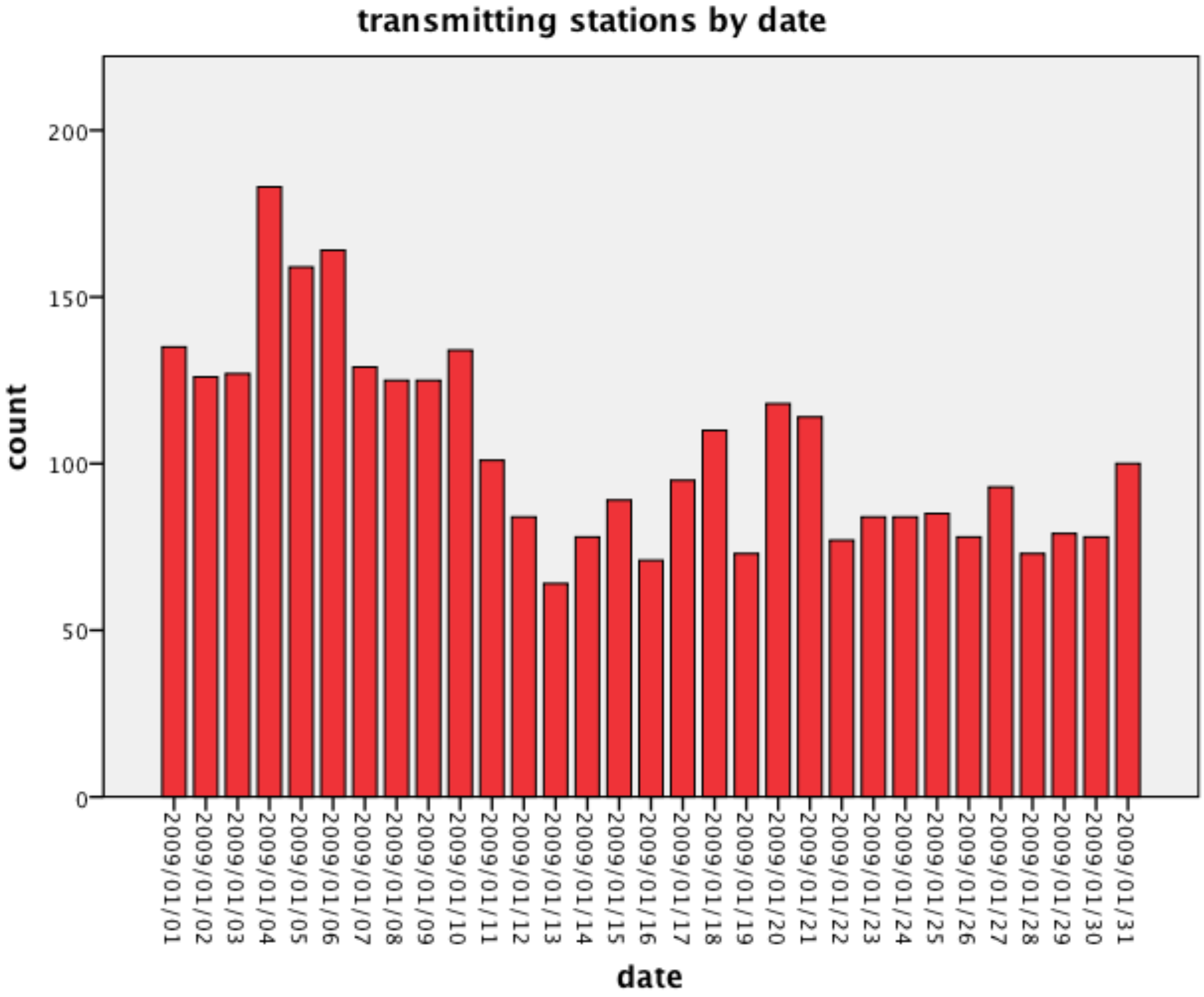
Did not significantly change the overall description of the data, and I now think that I should have chosen a higher minimum distance to work with.

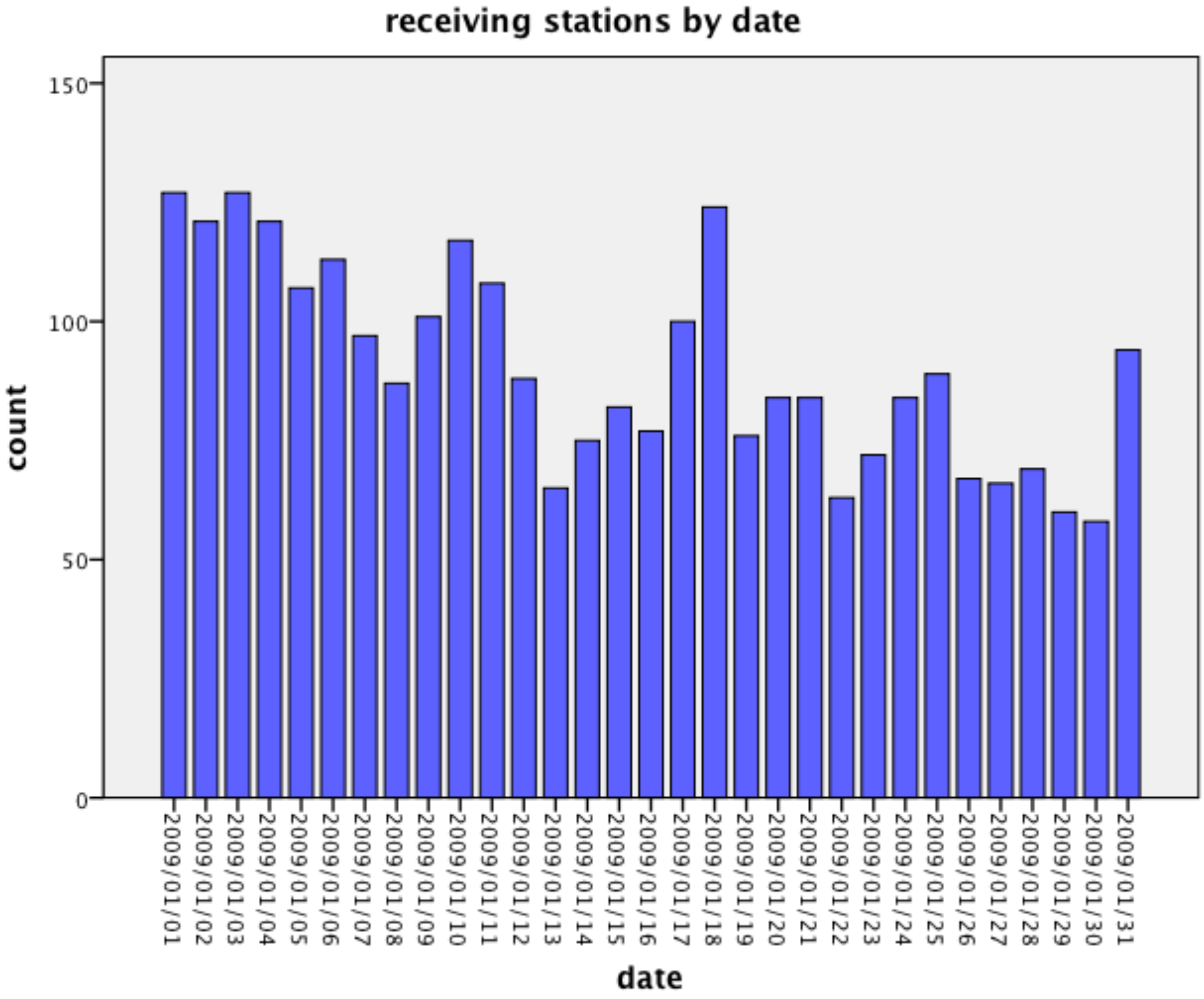
Step 12 – Looking at other data facets
While playing about with SPSS, I thought it might be interesting to see how many transmitting and receiving stations were making reports during the times that were being analyzed. I had the wspr spot data in a local MySQL database and was able to make an ODBC connection to it from the application. While I was able to do simple queries against the database, I was unable to work out how to do a more complex query. I ended up doing a query directly against the database, and exporting the data to a cvs file. This was more for playing about with SPSS, than anything else, other than seeing a small increase in the number of participating stations over the 4 year period, I don't thing anything meaningful can be obtained from the data.

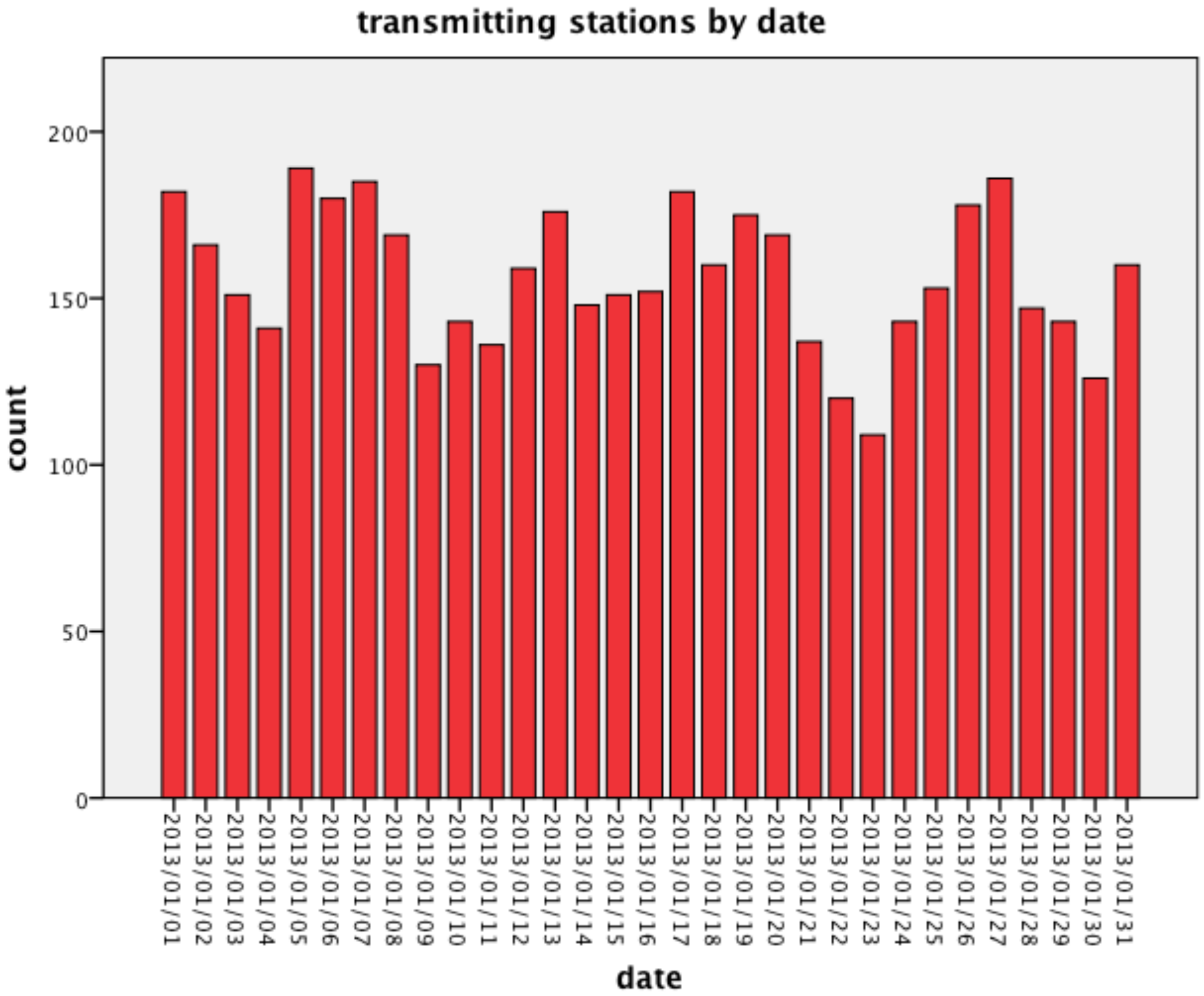
The basic format of the query was:

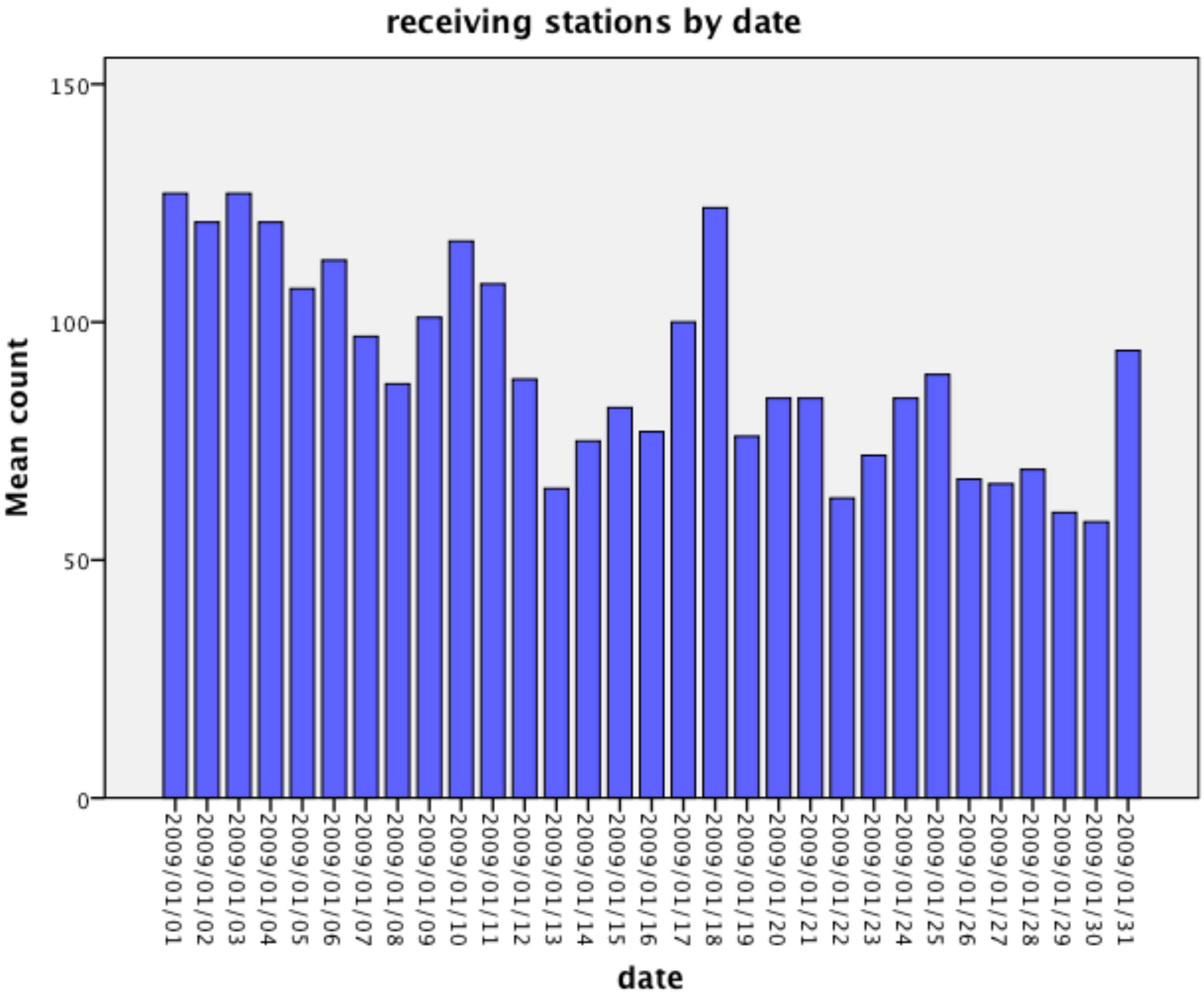
```
select count(*), date from (select distinct tx_station, date from
spots where date like '2009/01%' and band=10) tx_data
group by date
order by date;
```

Similar queries were exported for receiving stations by day, and for 2013.




receiving stations by date
150
100
50
0
count
2009/01/01
2009/01/02
2009/01/03
2009/01/04
2009/01/05
2009/01/06
2009/01/07
2009/01/08
2009/01/09
2009/01/10
2009/01/11
2009/01/12
2009/01/13
2009/01/14
2009/01/15
2009/01/16
2009/01/17
2009/01/18
2009/01/19
2009/01/20
2009/01/21
2009/01/22
2009/01/23
2009/01/24
2009/01/25
2009/01/26
2009/01/27
2009/01/28
2009/01/29
2009/01/30
2009/01/31
date


transmitting stations by date
count
200
150
100
50
0
2013/01/01
2013/01/02
2013/01/03
2013/01/04
2013/01/05
2013/01/06
2013/01/07
2013/01/08
2013/01/09
2013/01/10
2013/01/11
2013/01/12
2013/01/13
2013/01/14
2013/01/15
2013/01/16
2013/01/17
2013/01/18
2013/01/19
2013/01/20
2013/01/21
2013/01/22
2013/01/23
2013/01/24
2013/01/25
2013/01/26
2013/01/27
2013/01/28
2013/01/29
2013/01/30
2013/01/31
date



receiving stations by date

Alex V Flinsch
Data Analytics – Paper 2

Lessons Learned

While I was able to see some correlation between sunspot levels and mean distances of signal propagation, I was not able to show propagation paths as being open a bit longer than current models predict. This would have involved much more computation than I was prepared to do. The timestamp in the wspr data is in UTC, but local time (to determine day/night) for both stations involved in each case would need to have been calculated (based on lat/long determined from the corresponding grids). Once that had been done a percentage probability of a path being open between those points would need to be calculated, and then see if the actual path could have occurred at the time.

One thing that I did see is that the wspr data is somewhat dirty. While looking at the raw data I did see several cases where the reported band was outside of the allocated amateur radio frequency range. While this data may have been an actual reception report, I tend to doubt that it was. I have seen several of my signals reported outside of the band that I was transmitting on at the time, and when checking with the reporter, I was informed that it was an error on their part (wspr software not correctly connecting to their receiver's data port correctly).

Additional things learned –

- I don't remember as much about statistics as I thought I did. I will need to see if I can find some of my old text books and review them, or ask a family member involved in statistics if he has any that I can borrow.
- SPSS is way easier to use than it was in the past. I thing this analysis would have taken much longer if we were still doing data entry on 80 column cards. The data preparation alone would have taken me several months.
- SPSS seems to have limited ways to import dates, there does not seem to be a way to specify a user-defined format. This is unlike Excel, which converts anything that ***might*** be a date into a date. On additional reflection, the SPSS way may be better; Excel has burned me more than once with date conversions in the past.
- I will need to explore some additional features of SPSS. Specifically, I would like to know how to do more than a simple 'select a, b, c from x where…' statement. This would be useful in extracting data from existing sources. I would have tried to explore this method of data retrieval in depth, but was having difficulties in getting the ODBC connection set up correctly (the problem turned out to be an empty rogue .odbc.ini file).