

Parsing H5N1 Outbreak Data Using R & Python

Ryan Garner

May 29, 2009

Introduction

H5N1, also known as "bird flu", is a subtype of the Influenza A virus. The first recorded H5N1 outbreak occurred in 2004 in Southeast Asia. Since then, H5N1 has spread to countries such as China, Egypt, Indonesia, Pakistan, and Vietnam. Even though H5N1 currently appears to be contained in Asia through vaccinations, one might still question whether or not H5N1 could eventually spread to North and South America. Using spatial analysis, we can scientifically explore this question in order to better understand how the spread of H5N1 might play out in the future.

The future spread of H5N1 is currently being researched by Harry Kim. My contribution to Harry's research is to extract H5N1 outbreak data from PDF reports generated by the *World Health Organisation for Animal Health*. The parsed data can then be used for spatial analysis in **R**. Furthermore, in this paper I will discuss the tools I used to parse H5N1 outbreak data from PDF reports.

Data

The data to be extracted for spatial analysis in **R** is contained within PDF reports generated by the *World Health Organisation for Animal Health*. These reports are available at <http://www.oie.int>. Each report consists of multiple outbreak reports that occurred in different regions within a specified country.

Follow-up report No.: 11

Report reference: , OIE Ref: 6858, Report Date: 25/02/2008, Country: China (People's Rep. of)

Report Summary

Disease	Highly pathogenic avian influenza	Animal type	Terrestrial
Causal Agent	Highly pathogenic avian influenza virus	Serotype(s)	HSN1
Clinical Signs	Yes	Reason	Reoccurrence of a listed disease
Date of first confirmation of the event	19/04/2006	Date of Start of Event	13/04/2006
Date of report	25/02/2008	Date Submitted To OIE	25/02/2008
Diagnosis	Clinical, Laboratory (advanced), Necropsy	Date Of Last Occurrence	03/2006
Number Of Reported Outbreaks	Submitted= 14, Draft= 0	Name of Sender of the report	Mr Jia Youling
Address	Ministry of Agriculture BEIJING	Position	National Chief Veterinary Officer
Telephone	(86-10) 641 928 33 / 641 928 28	Fax	(86-10) 641 924 68 ou 928 69
Email	xmjwjch@agri.gov.cn	Entered by	Mme Maria Cristina Ramirez

Outbreak (other report - submitted)

LIAONING	Unit Type	Location	Latitude	Longitude	Start	End
LIAONING	Not applicable	Panjin	39	121	13/04/2006	13/04/2006
Species	Measuring units	Susceptible	Cases	Deaths	Destroyed	Slaughtered
Wild species	Animals	...	1	1	0	0
Affected Population		a magpie found dead				

Outbreak (other report - submitted)

LIAONING	Unit Type	Location	Latitude	Longitude	Start	End
LIAONING	Not applicable	Jinzhou	39,1	121,7	13/04/2006	13/04/2006
Species	Measuring units	Susceptible	Cases	Deaths	Destroyed	Slaughtered
Wild species	Animals	...	1	1	0	0
Affected Population		a wild duck found dead				

Outbreak (other report - submitted)

QINGHAI	Unit Type	Location	Latitude	Longitude	Start	End
QINGHAI	Not applicable	Yushu (county)	35,6065	95,8881	23/04/2006	
Species	Measuring units	Susceptible	Cases	Deaths	Destroyed	Slaughtered
Wild species	Animals	...	533	533	0	0

Figure 1: Sample outbreak report.

The data fields we are interested in extracting information from are: *Location, Latitude, Longitude, Start, End, Species, Measuring Units, Susceptible, Cases, Deaths, Destroyed, Slaughtered* and *Affected Population*.

Python PDFMiner

To parse the data in **R**, we first convert the PDF reports into HTML using PDFMiner – a **Python** package used to extract and analyze text data in PDF documents. PDFMiner can be downloaded from <http://www.unixuser.org/~euske/python/pdfminer/index.html>. Since the converted HTML documents retain the positioning of the data, we can utilize *regular expressions* to parse the required information.

R Parse

Once the PDF reports were converted to HTML, we then parse the required information using the **R** function I developed – `parse.R`. This function inputs an HTML file and outputs an **R** data frame containing the required information. To find the data in the HTML, we use pattern matching via *regular expressions*. As the function reads each line, it searches for the beginning of an outbreak section. Once it finds an outbreak section, it then proceeds to read until it finds the first required field – *Location*. From there, the positioning information of all the required fields is stored in memory and the data contained in each field is parsed based on this information. Since the positioning of the data is not static in each outbreak section (see Figure 1), this function takes into account a small range of where the data could lie in the HTML. By sampling a couple of HTML documents, I was able to find a suitable positioning range for each field. Also, as this function parses the data, it properly formats dates, latitudes, and longitudes along with converting empty fields to "NA". Finally, once all the data has been parsed from an outbreak section, a data frame is created and returned.

Conclusion

In this paper I discussed how I parsed H5N1 data from PDF files using such tools as **Python**, **R**, and *Regular Expressions*. I found this project to be quite fulfilling as I learned how to solve a difficult problem using multiple tools. All in all I ended up extracting approximately 11,000+ observations from 300+ PDF files. This data parsing task took my computer approximately 5-7 minutes to convert the PDF's and approximately 5-7 minutes to parse the data in R. One can only imagine how long something like this would take using a brute force copy and paste method.

Outbreak (other report - submitted)

LIAONING	Unit Type	Location	Latitude	Longitude	Start	End
LIAONING	Not applicable	Panjin	39	121	13/04/2006	13/04/2006
Species	Measuring units	Susceptible	Cases	Deaths	Destroyed	Slaughtered
Wild species	Animals	..	1	1	0	0
Affected Population		a magpie found dead				



pdf2text.py

```
<span style="position:absolute; writing-mode:lr-tb; left:31px; top:328px; font-size:9px;">Outbreak (other report - submitted) </span>
<span style="position:absolute; writing-mode:lr-tb; left:31px; top:344px; font-size:6px;">LIAONING</span>
<span style="position:absolute; writing-mode:lr-tb; left:139px; top:344px; font-size:6px;">Unit Type</span>
<span style="position:absolute; writing-mode:lr-tb; left:257px; top:344px; font-size:6px;">Location</span>
<span style="position:absolute; writing-mode:lr-tb; left:439px; top:344px; font-size:6px;">Latitude</span>
<span style="position:absolute; writing-mode:lr-tb; left:542px; top:344px; font-size:6px;">Longitude</span>
<span style="position:absolute; writing-mode:lr-tb; left:677px; top:344px; font-size:6px;">Start</span>
<span style="position:absolute; writing-mode:lr-tb; left:799px; top:344px; font-size:6px;">End</span>
<span style="position:absolute; writing-mode:lr-tb; left:31px; top:355px; font-size:6px;">LIAONING</span>
<span style="position:absolute; writing-mode:lr-tb; left:139px; top:355px; font-size:6px;">Not applicable</span>
<span style="position:absolute; writing-mode:lr-tb; left:257px; top:355px; font-size:6px;">Panjin</span>
<span style="position:absolute; writing-mode:lr-tb; left:456px; top:355px; font-size:6px;">39</span>
<span style="position:absolute; writing-mode:lr-tb; left:561px; top:355px; font-size:6px;">121</span>
<span style="position:absolute; writing-mode:lr-tb; left:577px; top:355px; font-size:6px;">13/04/2006</span>
<span style="position:absolute; writing-mode:lr-tb; left:696px; top:355px; font-size:6px;">13/04/2006</span>
<span style="position:absolute; writing-mode:lr-tb; left:31px; top:366px; font-size:6px;">Species</span>
<span style="position:absolute; writing-mode:lr-tb; left:144px; top:366px; font-size:6px;">Measuring units</span>
<span style="position:absolute; writing-mode:lr-tb; left:342px; top:366px; font-size:6px;">Susceptible</span>
<span style="position:absolute; writing-mode:lr-tb; left:463px; top:366px; font-size:6px;">Cases</span>
<span style="position:absolute; writing-mode:lr-tb; left:560px; top:366px; font-size:6px;">Deaths</span>
<span style="position:absolute; writing-mode:lr-tb; left:667px; top:366px; font-size:6px;">Destroyed</span>
<span style="position:absolute; writing-mode:lr-tb; left:776px; top:366px; font-size:6px;">Slaughtered</span>
<span style="position:absolute; writing-mode:lr-tb; left:31px; top:378px; font-size:6px;">Wild species</span>
<span style="position:absolute; writing-mode:lr-tb; left:144px; top:378px; font-size:6px;">Animals</span>
<span style="position:absolute; writing-mode:lr-tb; left:371px; top:378px; font-size:6px;">..</span>
<span style="position:absolute; writing-mode:lr-tb; left:478px; top:378px; font-size:6px;">1</span>
<span style="position:absolute; writing-mode:lr-tb; left:577px; top:378px; font-size:6px;">1</span>
<span style="position:absolute; writing-mode:lr-tb; left:693px; top:378px; font-size:6px;">0</span>
<span style="position:absolute; writing-mode:lr-tb; left:807px; top:378px; font-size:6px;">0</span>
<span style="position:absolute; writing-mode:lr-tb; left:31px; top:391px; font-size:6px;">Affected Population</span>
<span style="position:absolute; writing-mode:lr-tb; left:184px; top:391px; font-size:6px;">a magpie found dead</span>
```



parse.r

```
> head(h5n1)
  location latitude longitude start end species units susceptible cases deaths destroyed slaughtered affected
1 Panjin 39.0000 121.0000 2006-04-13 2006-04-13 Wild species Animals NA 1 1 0 0 a magpie found dead
```

Figure 2: H5N1 data parsing workflow.