Outbreaks

Mladen Cucak

5/4/2020

(Very relevant paper by Zwankhuizen & Zadoks)[<https://bsppjournals.onlinelibrary.wiley.com/doi/full/10.1046/j.1365-3059.2002.00738.x>]

I guess the main idea of this work is to determine the effect of between-season (winter) and pre season (spring) conditions on the number and rate of reported disease outbreaks.

## Data

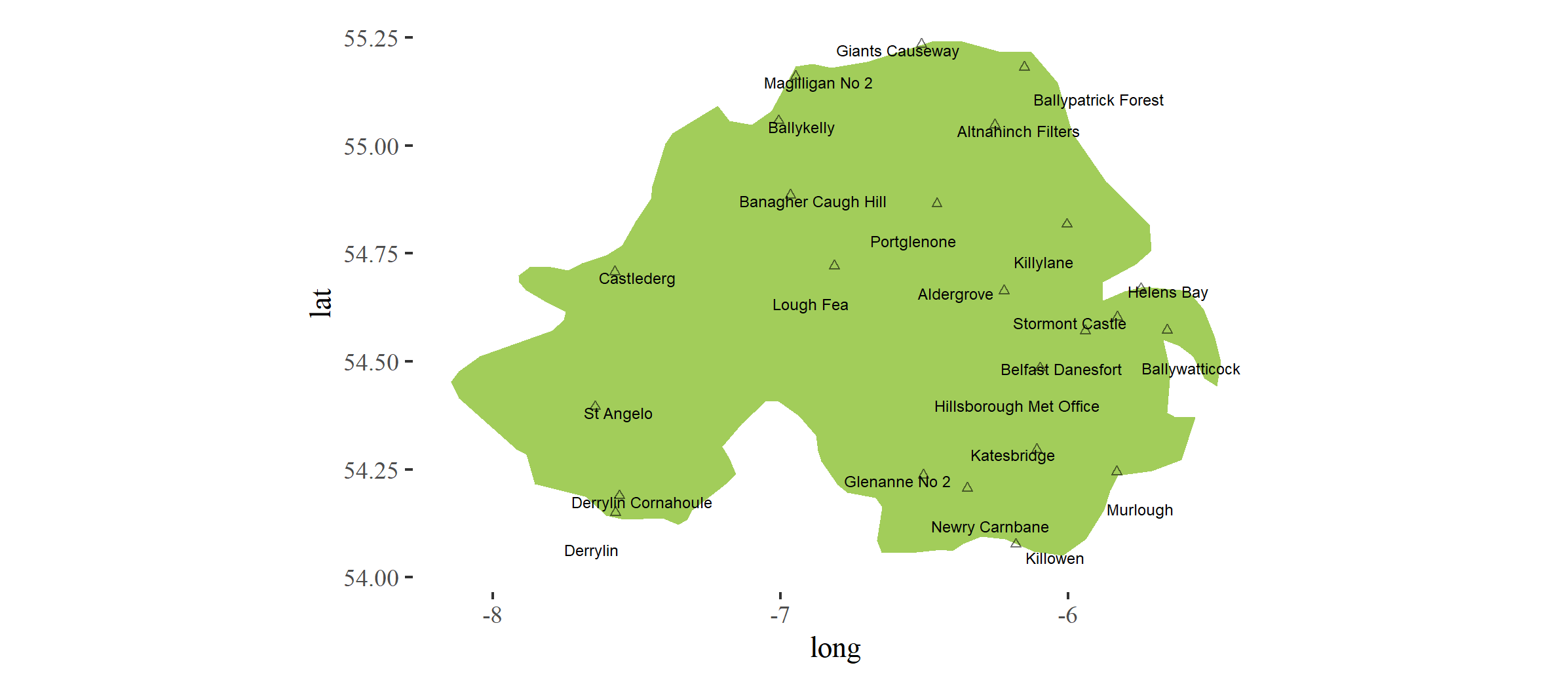
All this data relates to the basic data description so I will answer them in batch.

***1. Exactly what is on the vertical axes? [Maybe it is… data from ~200 sites in RoI (or on the island of Ireland?), each site scored for PLB on a presence/absence basis on a number of dates for a series of years?] Is the presentation of these graphs related in any way to that of graphs like Fig 1.7 in your BlightR manuscript, where the vertical axis is "Proportion of predicted outbreaks (TPR)"?***

Yes to all, this is the data used for validation of BlightR.

***5. Are there any weather/climate data available for all or part the corresponding time periods? At every site? For every year?***

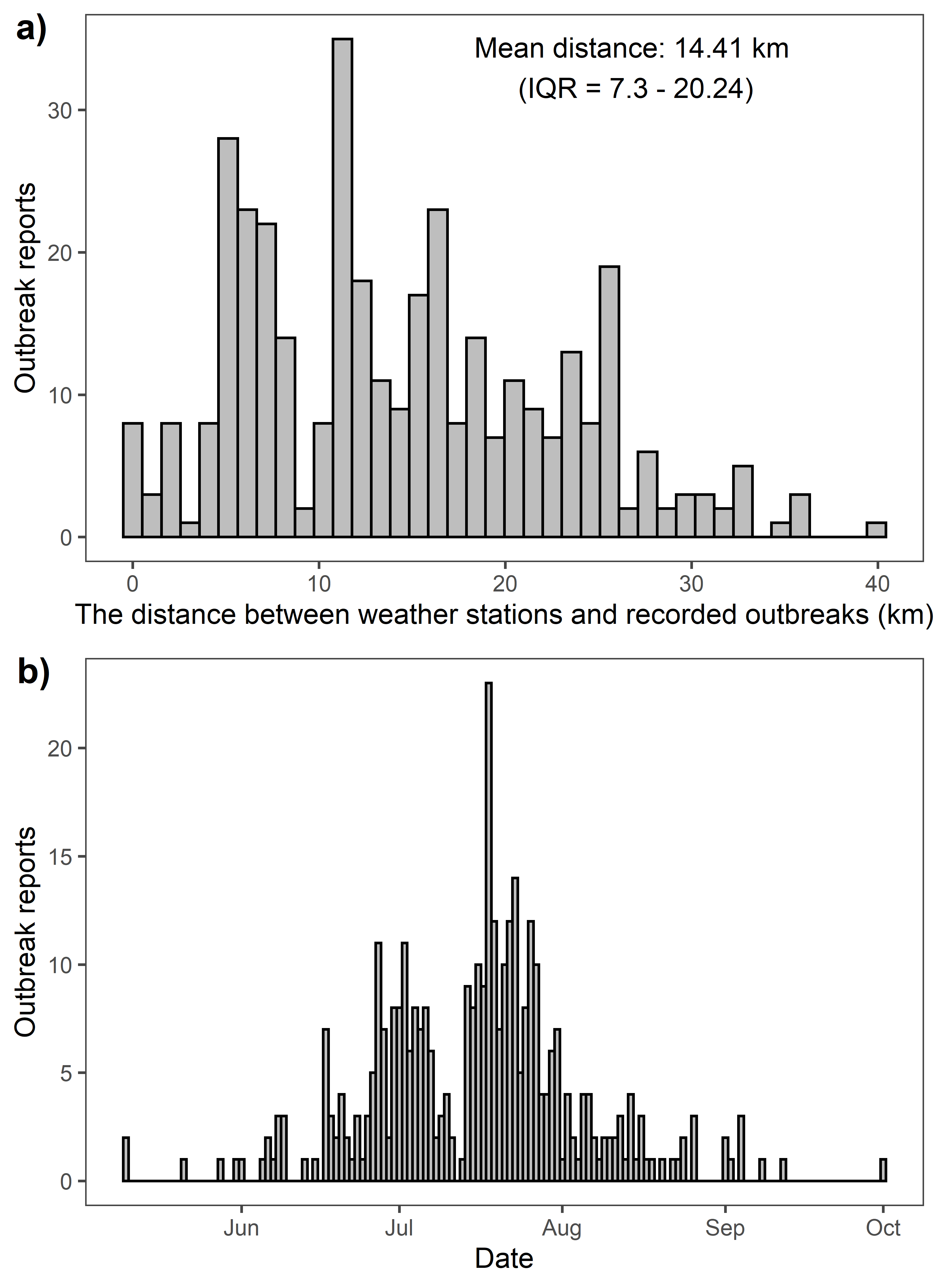
Yes, the weather data is available as described in BlightR. Briefly, the closest weather station, with less than 1% of missing data, was found iteratively and the weather data was assigned to each outbreak.



However, for this work, I had in mind calculating the synoptic averages (as per Zwankhuizen and Zadox, and Neil’s previous work) and use that data for “explaining” the model parameters. Matter a factm they have used a sigle station in centre of Netherlands to explain level of blight in following year which I find very questionable.

*“That's great, but we're now talking about two different kinds of model. Fitting a cdf or some other kind of nonlinear function to the data as a function of time would be a simple regression exercise where we'd estimate (probably at most 4) a few parameters. The choice of function should be motivated so that those parameters can be described in terms of biological and physical processes, and then a more mechanistic (or at least less phenomenological) approach would follow in which the parameters would be "unpacked" in terms of weather, or other, data.”*

Such averages could be much more than what they have done.



Bar charts showing the number of outbreak reports per: a) the distances up to 40 km from weather stations; and b) Julian day of year. (as in BlightR)

1. ***NEIL: Did you already make "predictions" for these data based on historical data sets (this pre-supposes that the answer to Gareth's Q5 is "yes")***

No, the data was used in different manner for BlightR. I simply used 10 days of weather prior to the outbreak to see if there were conditions for individual outbreaks.

***2. Was it the same sites in every year?***

No, see more below.

***3. Was each site a single crop, or do (some or all) data points represent some kind of multi-crop regional assessment?***

No, see more below.

***4. Were the assessment dates systematic in any way (e.g., by time, or by crop growth, etc.)?***

No, see more below.

***6. What about other (e.g., agronomic, geographic) covariate data for the sites / crops (that might be relevant to PLB)?***

Yes, see more below.

***7. Was PLB presence / absence determined by visual assessment? Was there a pre-specified sampling/surveillance scheme? Were the assessments done in the same way at each site?***

The data originates from Norther Ireland and was collected as part of the Agriculture and Horticulture Development Board (AHDB) Potatoes ‘Fight Against Blight’ campaign (<https://potatoes.ahdb.org.uk/>). It is consisted of a single disease outbreak report from Blight Scouts (people that volunteer to walk crops and are familiar with sampling procedure), or simply from producers sending samples for testing.

The dataset is consisted of the date and coordinates of 352 late blight outbreaks from across Northern Ireland over an 11- year period (2005-2014). The geographic coordinates of blight outbreak locations were obtained programmatically (Because I had only the name of the village, barony and county) and confirmed manually (using google maps and open street maps). Most of the data was from field crops (conventional production). There is no data about the disease level.

A look at data set.

date: Date when the outbreak was observed.

yr: Year

source: type of production

variety: variety that was grown

variety: sometimes there are multiple varieties (when I was getting samples in the lab, if two varieties are reported it means the disease was wide spread in the field).

Location and county: Exact name of village, barony (or another level administrative unit) and county.

comments: Occasionally there were some comments such as protected crop, or number of samples from that field, etc…

Loc\_full: Entire location reference I used for programmatic search of geographical coordinates.

lat and lon: are variables I have calculated

jday: extracted from the date

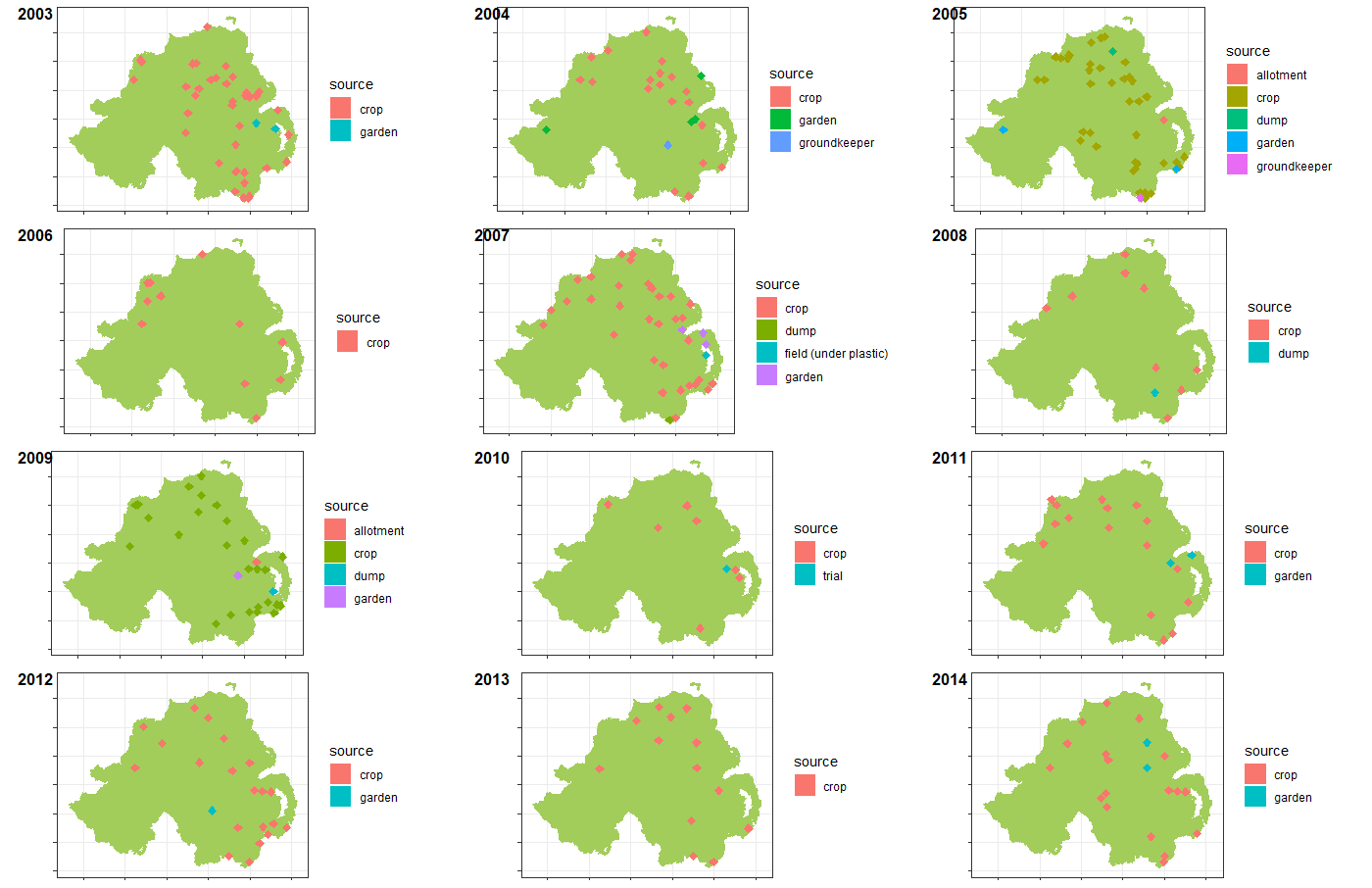
Table continues below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| date | yr | source | variety | varietyII | Location |
| 2003-05-21 | 2003 | crop | Lady Felicia | NA | Ards |
| 2003-05-31 | 2003 | crop | Home Guard | NA | BallyLagen,BClare |
| 2003-06-05 | 2003 | crop | Kerrs Pink | NA | Benagh, Kilkeel |
| 2003-06-06 | 2003 | crop | Kerrs Pink | NA | Glenkeen Colraine |
| 2003-06-06 | 2003 | crop | Home Guard | NA | Pollee B.Mena |
| 2003-06-07 | 2003 | crop | Kerrs Pink | NA | Dunamoy B.Clare |

Table continues below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| County | Comments | loc\_full | lon | lat |
| Down | NA | Ards, Down, UK | -5.664 | 54.66 |
| Antrim | NA | BallyLagen,BClare, Antrim, UK | -5.922 | 54.76 |
| Down | protected crop | Benagh, Kilkeel, Down, UK | -6.078 | 54.05 |
| Londonderry | NA | Glenkeen Colraine, Londonderry, UK | -7.305 | 54.99 |
| Antrim | NA | Pollee B.Mena, Antrim, UK | -6.121 | 54.55 |
| Antrim | NA | Dunamoy B.Clare, Antrim, UK | -6.041 | 54.78 |
| jday |
| 141 |
| 151 |
| 156 |
| 157 |
| 157 |
| 158 |

Below is the map f blight outbreaks (apologies for the high resolution figure, need to work on it). Location of each outbreak is plotted for each year. Different colors indicate the different source (as indicated in the legend), whether a grown crop, garden crop, allotments, etc.

 `

A lot of different varieties is present in the data. I know most of them, and all are traditional varieties without disease resistance.

## [1] "Lady Felicia" "Home Guard"   
## [3] "Kerrs Pink" "British Queen"   
## [5] "Desiree" "Saxon"   
## [7] NA "Up to Date"   
## [9] "Dunbar Standard" "D.Standard"   
## [11] "King Edward" "Dundrod"   
## [13] "Sante" "Nicola"   
## [15] "Red Duke of York" "Navan"   
## [17] "Rooster" "Pentland Squire"   
## [19] "Arran Victory" "Milagro"   
## [21] "Sharpes Express" "Arran Banner"   
## [23] "Caberet" "Dunluce"   
## [25] "Maris Piper" "Desiree & Nicola"   
## [27] "Lady Claire" "Golden Wonder"   
## [29] "Pentland Ivory" "Marfona"   
## [31] "Pentland Javelin" "Pink Fir Apple"   
## [33] "Avalanche" "UpToDate"   
## [35] "Santee" "Romano"   
## [37] "Piccolo Star" "various"   
## [39] "Charlotte" "Lady Rosetta"   
## [41] "Paramount" "R1"   
## [43] "Up To Date" "Cara"   
## [45] "Sarpo Mira" "Hermes"   
## [47] "Cabaret" "Duke of York"   
## [49] "Tomato Money Maker" "Tomato Gardeners Delight"  
## [51] "Up-to-Date" "AFLO1\_1"   
## [53] "VR808" "Golden Nugget"   
## [55] "Aphrodite" "Kifli"   
## [57] "Ramos" "Up-To-Date"

This is the data summary for different sources per each year.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| source | | 2003 | | 2004 | | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
| allotment | | 0 | | 0 | | 2 | 0 | 0 | 0 | 1 | 0 |
| crop | | 43 | | 21 | | 46 | 15 | 43 | 9 | 36 | 9 |
| dump | | 0 | | 0 | | 1 | 0 | 1 | 1 | 1 | 0 |
| field (under plastic) | | 0 | | 0 | | 0 | 0 | 1 | 0 | 0 | 0 |
| garden | | 2 | | 4 | | 2 | 0 | 3 | 0 | 1 | 0 |
| groundkeeper | | 0 | | 1 | | 1 | 0 | 0 | 0 | 0 | 0 |
| trial | | 0 | | 0 | | 0 | 0 | 0 | 0 | 0 | 1 |
| 2011 | 2012 | 2013 | 2014 | |
| 0 | 0 | 0 | 0 | |
| 25 | 35 | 16 | 32 | |
| 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | |
| 2 | 1 | 0 | 2 | |
| 0 | 0 | 0 | 0 | |
| 0 | 0 | 0 | 0 | |

***8. Once a site was scored “PLB present”, did assessments continue, or was that it (for that site). Or could a “PLB present” site subsequently be scored “PLB absent”?***

Most of this data is probably collected on the bases of someone seeing blight in the crop and senting the sample. Dates in the dataset are dates when the disease was observed.

Only thing I can say with some level of certainty is that there was probably no disease before the outbreak, however there are always other possibilities.

A possible scenario is: Blight was present for a couple of weeks (or even a month depending on the weather conditions) before the report: A grower gets blight, applies pesticide, and it does not work and blight re-appears in other part of the field, which is when the grower decides to send samples for testing. These would be outbreaks caused by the resistant strains.

Blight is a high risk disease and as soon as producers see some blight they would apply systemic fungicides and burn of the patch with the infection in radius of >2 m. Hence there are no follow-up reports.

***9. I take it that these were all treated crops? If so, were they all treated according to the same PLB “early warning” system, or was a variety of such systems in use (or none)? Was this information recorded, crop-by-crop? Do we know the dates for start and finish of treatments, relative to the start and finish of the disease assessments?***

I believe the information provided above answers this question. However, if it was a field crop, it I safe to assume that all of them were treated.

***10. Are the fitted curves based on binary logistic regression analysis; date (continuous) on the horizontal axis and then at each date, a set of 0s & 1s representing the recorded absence / presence data (the data point for each date on the graph being the total of 1s)? Did the 2008 data just “fail”?***

As described above, there is no “absence” data. We only have “presence” data.

1. ***Neil: I'd like a bit more detail on what Rafael did to fit the logistic function to those different datasets.***

Initially, a simple cumulative sum of outbreaks in each year was calculated.

Yr – year (1st year is 2003) (grouping factor)

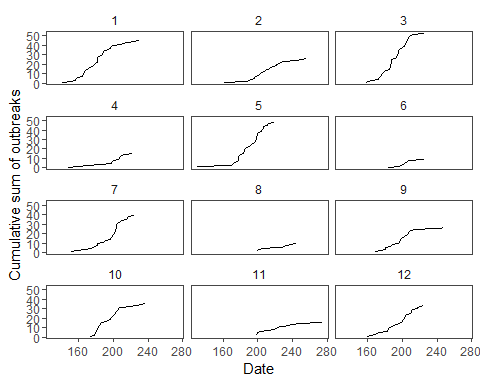
Jday – day of the year (continuous predictor)

Cums – Cumulative sum of outbreaks (response)

A look the piece of resulting data set to get an idea:

|  |  |  |
| --- | --- | --- |
| yr | jday | cums |
| 1 | 141 | 1 |
| 1 | 151 | 2 |
| 1 | 156 | 3 |
| 1 | 157 | 5 |
| 1 | 158 | 6 |
| 1 | 164 | 7 |
| 1 | 168 | 13 |
| 1 | 170 | 14 |
| 1 | 171 | 16 |
| 1 | 177 | 19 |
| 1 | 178 | 20 |
| 1 | 179 | 21 |
| 1 | 181 | 22 |
| 1 | 182 | 27 |
| 1 | 184 | 28 |
| 1 | 186 | 29 |
| 1 | 188 | 32 |
| 1 | 189 | 33 |
| 1 | 191 | 34 |
| 1 | 197 | 37 |
|  |  |  |

To visualize the entire data set:



## Model fitting

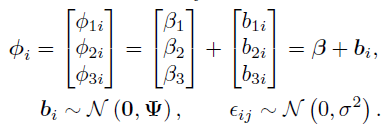
Logistic nonlinear mixed effects model was fitted using maximum likelihood estimation method.

This model expresses the cumulative sum of disease outbreaks yij per each day of year xij for i = 1, . . . , 52 and j = 1, . . . ,11 as

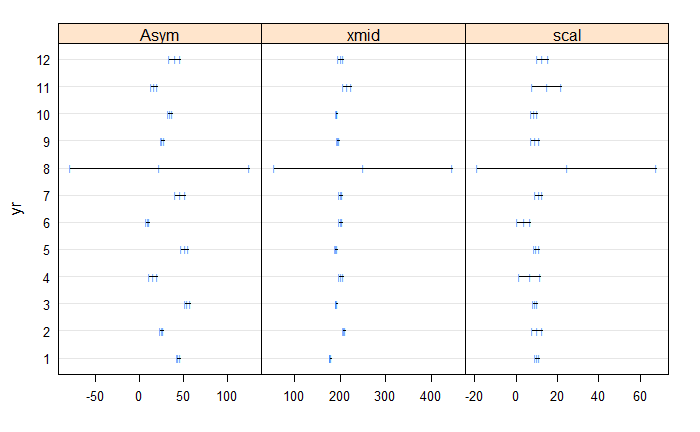


where the error terms Yij are assumed to be distributed independently as N(0, σ2). the model parameters are the asymptotic cumulative sum of disease outbreaks φ1, the day of the year at which the half of its asymptotic cumulative sum of disease outbreaks is attained φ2, and the growth scale (rate) φ3. This function is nonlinear in φ2 and φ3.

The model parameters are allowed to vary with year



Year 8 was removed because the data was no sufficient to fit the logistic model.



***Did you calculate any g-o-f statistics (I’m not looking for numbers, just interested which ones, if any)?***

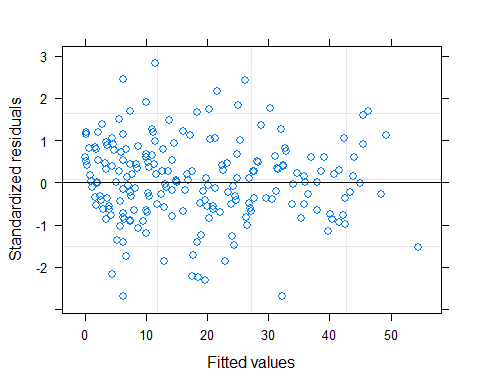
No, goodness of fit for mixed effect models are veru questionable, so GOF was simply explored using diagnostic plots.

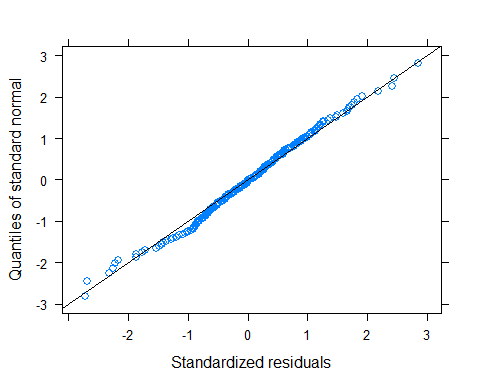
The two basic distributional assumptions:

Assumption 1 : the within-group errors are independent and identically normally distributed, with mean zero and variance σ2, and they are independent of the random effects.

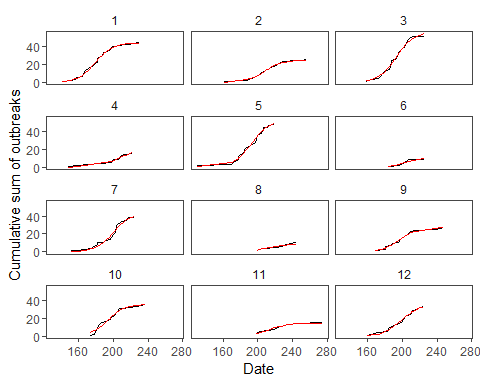
Assumption 2 : the random effects are normally distributed, with mean zero and covariance matrix Ψ (not depending on the group) and are independent for different groups;

A plot of the standardized residuals versus the fitted values shows that the residuals are distributed symmetrically around zero, with an approximately constant variance. It does not indicate any violations of the assumptions for the within-group error.





The adequacy of the fitted model is better visualized by displaying the fitted and observed values in the same plot.



1. ***Neil: Have you tried any sort of cluster or pattern analysis with these? For example (and just to start the conversation) it looks like years 1, 3 and 5 are similar, with 2,4,6 and 11 also being a loose group and 7.9.10 and 12 a third group?***

I did not do anything of the sort just yet. That could be the next step. However this classification variable does not necessarily need to be categorical but perhaps continuous because I am not sure if it would be fair to make 1 or 2 groups. However, there is certainly a lot of difference years you have mentioned here.

Perhaps some classification model could be fitted to