Two steps at a time

a very simple bigram model of cattle movement in Great Britain

Jess Enright and Rowland Kao
The Boyd Orr Centre for Population and Ecosystem Health, University of Glasgow

What's a bigram?

A bigram is a pair of tokens taken in order from asequence of tokens. We can use frequency counts of bigrams to model datasets.

application:

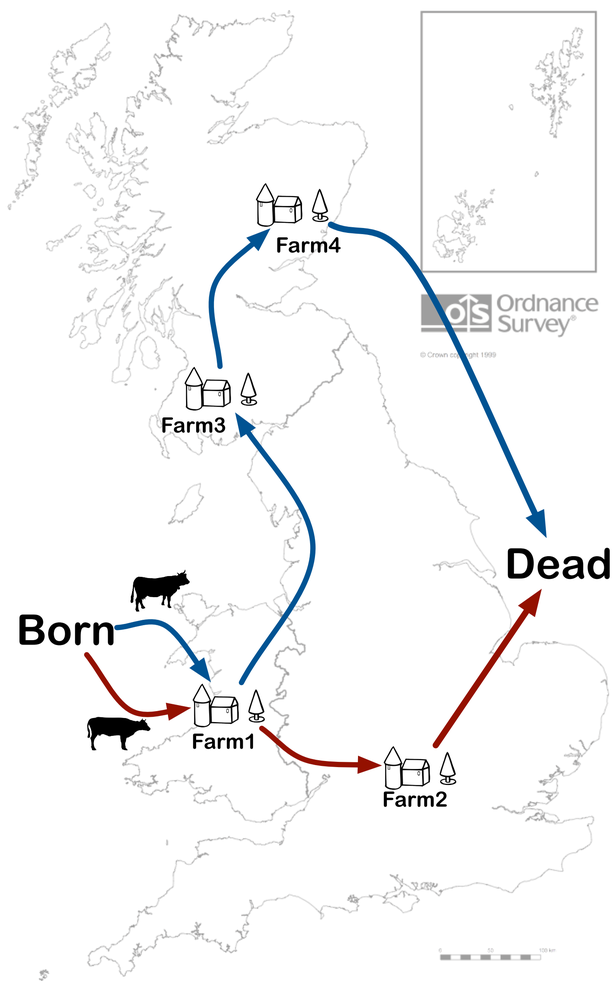
Look at the two cattle life histories on the right.

There are two cows. One has the life history:
Born, Farm1, Farm2, Dead
The other has:
Born, Farm1, Farm3, Farm4, Dead

The first life history contains three bigrams:
[Born, Farm1] [Farm1, Farm2] and [Farm2, Dead]

We can count the number of each bigram that occurs in these two life histories, producing the table on the far right.

Cattle movements in Great Britain are recorded by the British Cattle Movement Service (BCMS), so we have a very large number of life histories available. We have counted bigram frequencies over the BCMS dataset to build a simple model of that dataset.



Bigram	Count
Born Farm1	2
Farm1 Farm2	1
Farm1 Farm3	1
Farm2 Dead	1
Farm3 Farm4	1
Farm4 Dead	1

How can we use a bigram model?

for prediction:

If we had only partial life histories, or wanted to generate large simulated animal movement datasets, we could use bigram frequencies to produce predicted life histories.

We split cattle movements from 2001 to 2011 in two halves. We counted the bigrams in one to make a bigram frequency model, and then truncated a random selection of the other to form a test set. We used the model to predict the missing parts of the test set.

61% of the predictions made by the bigram model were correct.

Why restrict ourselves to bigrams? Trigrams are analogous to bigrams, but are of length three. We ran the same training and testing regime with a trigram model.

72% of the predictions made by the trigram model were correct.

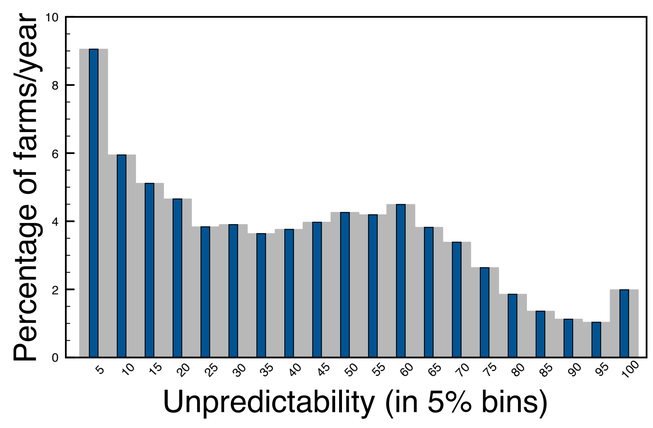
There are weaknesses:
- Neither model is as good at predicting long life histories
- These long life histories may be important in disease spread

for measuring unpredictability:

If we look at the change in bigram frequencies over time, we get an idea of the consistency of trade patterns over time.

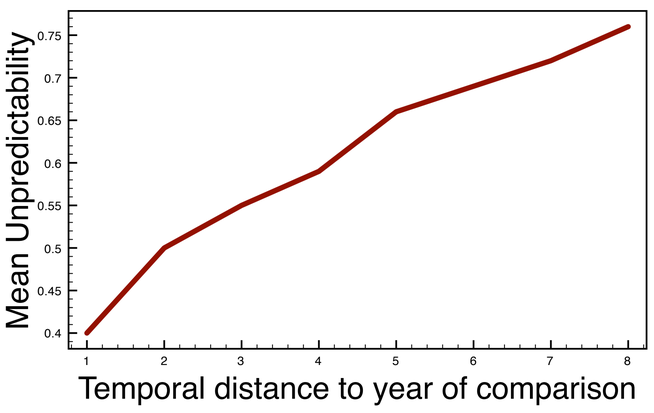
We computed the percentage of animals from each farm that are sent to a location that farm did not send animals to in a previous year - we call this the farm's **unpredictability.** This is the same as the percentage of bigrams starting at a farm that do not occur in that previous year.

We can look at the unpredictability of a year relative to the year immediately before it, or to years farther in the past.



What is the distribution of farm unpredictabilities relative to the previous year?

To the left, we've plotted discretized frequencies of unpredictabilities relative to the previous year over 2001 to 2010, considering only farms that have moved at least 50 animals.
We find that:
- low unpredictabilities are the most common
- half of the farms/year have unpredictability over 40%
- there is a small increase in frequency for highest 5% of unpredictability



Do trading patterns change over time?

To the left we've plotted the mean unpredictability of farms in 2010 relative to years 1, 2, 3, 4, 5, 6, 7, and 8 years in the past. Unpredictability increases with temporal distance. We would expect this if most farms gradually changed their movement patterns over time.

Farm movements in a year are more predicted by movements in recent years than to movements in years farther in the past.

acknowledgements

We gratefully acknowledge the support of the Scottish Government and the Centre of Expertise on Animal Disease Outbreaks, the Wellcome Trust, and thank Defra, RADAR, and the BCMS for cattle movement data.

jessica.enright@glasgow.ac.uk