

Cattle movements and bovine tuberculosis in Great Britain

Gilbert, M., A. Mitchell, D. Bourn, J. Mawdsley, R. Clifton-Hadley & W. Wint

Supplementary information

Supplementary information includes details of analyses and model developments, animated display complementing the paper, and complementary elements of discussion. The first section describes the variables and the multiple logistic regressions analysis of BTB distribution in 2002 and 2003 highlighting the role of cattle movement in annual predictions of BTB occurrences. The second section describes the methods used to build a multi-annual logistic model of BTB presence, and how this model is used to build a simulation model allowing short-term predictions. The last section provides complementary discussion points.

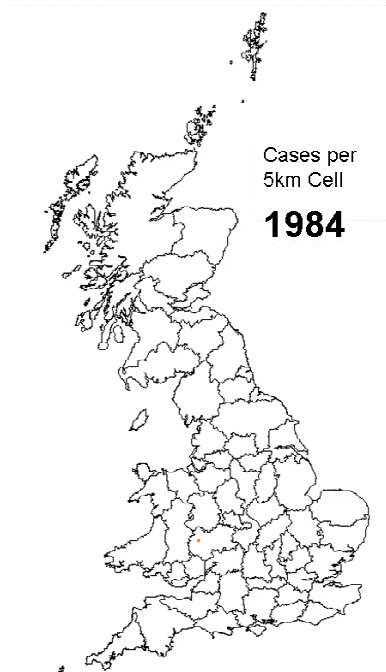
I Multiple logistic regression of BTB presence in 2002 and 2003.

The aim of this analysis was to demonstrate that animal movement data outperform other variables previously found to predict BTB distribution¹. This analysis thus involved the same data source, complemented by animal movement data only, and involved the same analysis procedure (data selection and variable entry rules) as in the previous analysis, so that changes in the results only reflects the incorporation of the animal movement data.

I.1 Data on BTB

Data on BTB presence were derived from the Defra VETNET database for the period 1984-2003 (Supplementary video 1). These data include BTB monitoring data for Great Britain, covering in excess of 90,000 holdings.

Spread of BTB in Britain: 1984-2003



Supplementary video 1. Animation showing the distributions of number of BTB detected cases per 5 km cell between 1984 and 2003 (open the Supplementary video 1 to display the animation).

1.2 Predictors

The following data sources are similar to those previously used in Wint *et al.*¹ to predict the distribution of BTB in 1997.

Land use and land cover. A variety of land use and land cover data was extracted as raster images from the Countryside Information System CD-ROM². These data record the percentage cover of 17 land use classes at 1 km grid cell resolution, simplified from data with 25 classes on a 25m grid, with a minimum unit of 0.125 ha. They result from a statistical classification of reflectance data from 46 Landsat TM scenes and the overall map accuracy is cited to be 80-85%. The following land use types were converted to 1 km resolution data layers as percentages (equivalent to ha/km): tilled area; deciduous woodland; coniferous woodland; managed grassland; urban land; suburban land; bare ground; and water.

Climate data. Remotely sensed data were derived from daily 1 km resolution imagery of the Advanced Very High Resolution Radiometer (AVHRR) on the National Oceanographic and Atmospheric Administration satellite and processed by the Pathfinder program^{3,4} to remove cloud and other atmospheric contamination. The following measures of atmospheric and land-surface characteristics, derived from multi-temporal satellite sensor imagery⁵ were used:

- 1) Normalised Difference Vegetation Index (NDVI) from the Advanced Very High Resolution Radiometer (AVHRR) commonly used as an indicator of vegetation cover (data from the Pathfinder Program);
- 2) A measure of land surface temperature (LST), derived using the Price split window technique from the thermal channels 4 and 5 of the AVHRR⁶;
- 3) A measure of Middle Infrared Reflectance (MIR), allied to temperature, but less susceptible to atmospheric interference, derived from Channel 3 of the AVHRR;
- 4) Vapour Pressure Deficit (VPD) from AVHRR Channels 4 and 5 and ancillary meteorological data;
- 5) Air temperature (AT) estimates, also derived from AVHRR satellite channels.

Data are available only within the period 1992/3–1995/6, and so were combined into monthly averages to provide complete temporal coverage of a nominal calendar year, then further processed to produce variables additional to the original imagery (Supplementary Table 1) using the algorithms described in Hay and Lennon⁵. The monthly data were subjected to temporal Fourier processing, previously shown to provide descriptive and explanatory variables associated with distributions of vectors and diseases. These describe the seasonal cycle in terms of sinusoidal annual, bi-annual and tri-annual components, each with an amplitude and phase⁷ (i.e. timing of the first peak). Additional data layers were produced showing the Fourier-fitted (i.e. essentially smoothed) maximum and minimum signal value and the contribution of each of the annual, bi-annual and tri-annual cycles to the overall variance of the seasonal signal.

Anthropogenic and demographic data. Human population data were derived from several sources: a global coverage of population number per image pixel, obtained from University of California at Berkeley provided by FAO AGL at 5 minute resolution; a population density coverage at the same resolution from the Consortium for International Earth Science Information Network (CIESIN: <http://www.ciesin.org>), derived from data collated by the National Centre for Geographic Information and Analysis (NCGIA: <http://www.ncgia.ucsb.edu>). A range of population related data was also extracted from the Landsat 1km resolution archive, including night-time light intensity, roads, each of which was recoded to presence and absence, from which distance to roads and distance to lights images were constructed (http://www.ornl.gov/gist/projects/LandScan/landscan_doc.htm).

Other Eco-climatic and Land Related Data

Topographic Digital Elevation Model (DEM) data were obtained from the global GTOPO30 1km resolution elevation surface, produced by the Global Land Information System (GLIS) of the United States Geological Survey, Earth Resources Observation Systems (USGS, EROS) data centre. River courses were obtained from the USGS EROS data centre HYDRO1k data archive at <http://edcdaac.usgs.gov/gtopo30/hydro/>, from which a “distance to rivers” image was prepared. Potential Evapo-transpiration (PET): mean, minimum and maximum decadal values were calculated from 1961-1990 averaged obtained from the Food and Agriculture Organisation (FAO) of the United Nations and re-sampled to a 0.01 degree resolution.

Cattle data A range of different cattle density figures was obtained for England Scotland and Wales – aggregated both by administrative areas of various types, and by 5 or 20 km grid. Main holding level, animal census data were provided for several years (1987, 1992, 1996, 1997, 1999, 2000, 2001, 2002) for England, together with recent parish level data for Scotland and small area data for Wales. These had been initially screened to prevent disclosure – i.e. figures were removed if they could be used to identify animal numbers in individual holdings. Full data for Scotland and England was released in due course, though not for Wales for which the most recent complete data available was for 1999. In order to avoid potential mismatches between disclosed and non-disclosed data, the combined administrative unit data for Scotland and Wales and comparatively coarse gridded data for England in 1999 were used as a general measure of cattle density. Herd size, and proportion of dairy cattle were available for 1995 and 1997 only, and were derived from the full BTB datasets of tested holdings (as distinct from the agricultural census data) used in previous work¹. Holding density was also extracted for these years, and used as a surrogate for herd density

Badgers. Detailed information about badger distribution in Great Britain has not been possible to obtain. Information published in summary form for 1988 and 1997⁸ was unavailable because the original data “were collected by volunteers on the strict understanding that they would remain confidential.” Alternative, freely available badger data are, at best, patchy. The geographically most complete source from the Countryside Information System, contains 1km resolution information on badger distribution from the Mammal Society, the Biological Records Centre and the British Deer Society surveys between 1965 and 1990⁹. Even when aggregated to 10km, these records appear unlikely to provide a very realistic representation of actual distribution. An additional GIS derived variable was also calculated to provide a continuous (as opposed to binary) predictor variable based on the distance to the nearest kilometre resolution reported presence.

The following data sources were used for the first time in this analysis:

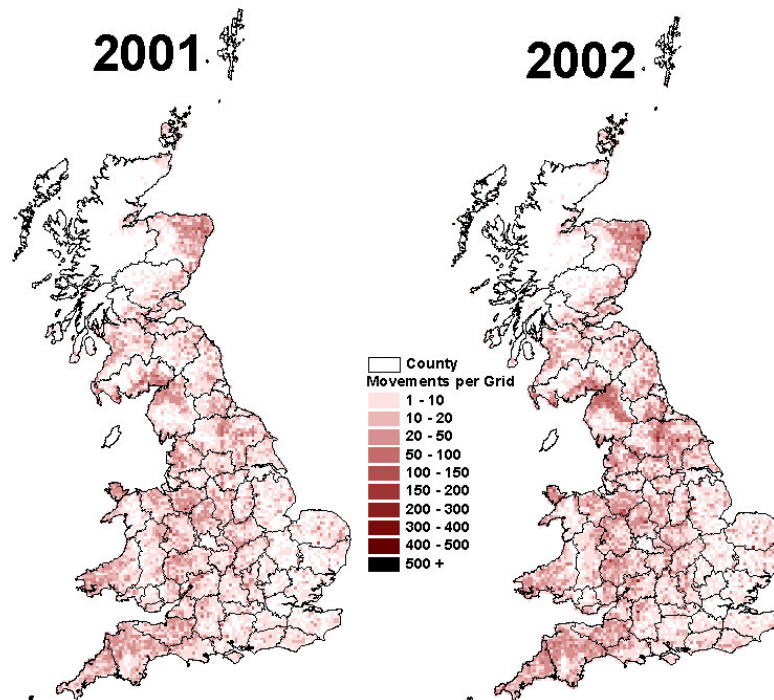
Animal movement data. The treatment of animal movement data extracted from the CTS involved three steps: geo-referencing movement locations, pairing “on” and “off” movements, and querying the obtained data set such as to obtain the variables required by the analysis.

The geo-referencing of movement locations used several methods in a hierarchical sequence. In the first instance, data held on the CTS (e.g. address and ordnance survey (OS) reference) were used to establish geographical coordinates (73.2% locations were identified using this method). If this failed, location data were cross-referenced against other data sources (e.g. VetNet, Census and slaughterhouse list, + 3.2%). Seventy-six percent of all locations and 98% of those associated with cattle movement were successfully geo-referenced.

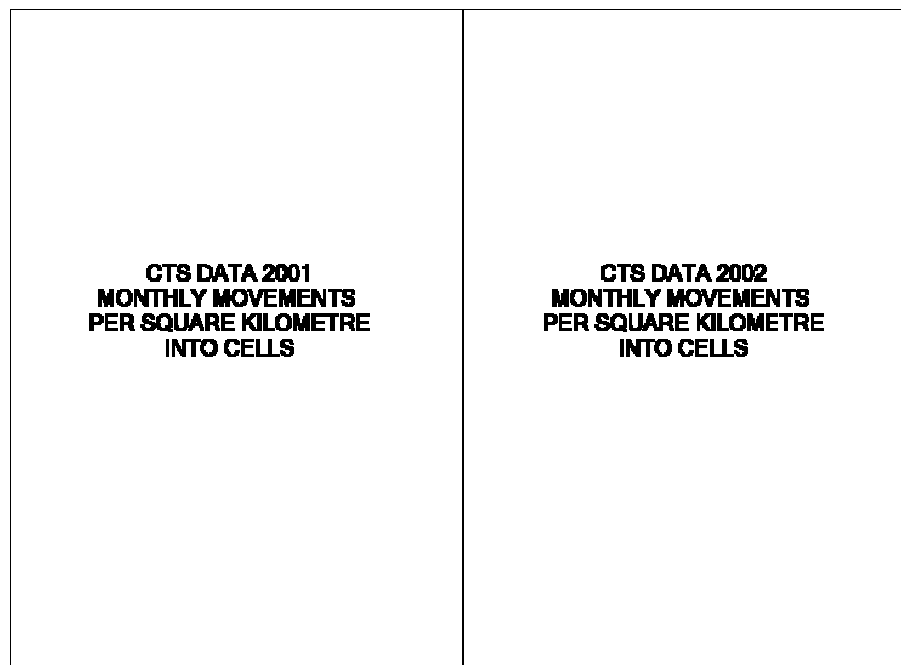
For all events, except births and deaths, the BCMS receives two records: one for the location the animal has moved off and the other record for the location the animal has moved onto. For any meaningful analysis of movement patterns, “off” and “on” records have to be “paired.” This process resulted in unresolved movement histories leading to the rejection of some animal’s movement records, in particular in the earlier years of the CTS (this procedure is detailed at length in Wint et al¹⁰, and in Mitchell et al.¹¹). However, this significantly improved in the recent years with 79%, 70%, 75% and 88% of animals born in the years 2000, 2001, 2002 and 2003 respectively having logical movement histories.

The paired movement database from the period 2000-2003 was then queried to estimate the following variables in each square: total number of inward movements, number of movements from infected areas, and the proportion of movements from infected areas, and these layers were added to the series of predictors obtained for each 1 km cell (see Supplementary Table 1).

The movement data for 2001 and 2003 and their monthly dynamics are illustrated in Supplementary Figure 1 and Supplementary Videos 2 & 3.



Supplementary Figure 1 Distribution of total ON cattle movements in 2001 and 2002



Supplementary Videos 2 & 3 Animations of monthly cattle movements for 2001 and 2002, which illustrate the virtual cessation of movements in March and April 2001, immediately after the outbreak of Foot and Mouth Disease (FMD) (open the supplementary video 2 & 3 to display the animations)

BTB persistence and spread. Some data were estimated from the distribution of BTB in the previous years such as to quantify BTB persistence and short-distance spread. These data included the number of years of past BTB infection in the 5km cell; number of infected cells in the previous year in a 5 km radius doughnut window; Distance to the nearest cell with BTB present.

All data variables and their abbreviations are presented in the Supplementary Table 1 below.

Supplementary Table 1 Predictor data used in the analysis of BTB occurrence

Type	Variable
Anthropogenic	Distance to roads (DROADS) and city lights (LIGHT)
Demographic	Human population density (HPOP)
Land cover	Percentage cultivation and managed grassland (PCU); Proportion urban and suburban land cover (PUR); Woodland (PWO), Open Woodland (POW); Grassland (PGR); Water (PWA); Normalised Deviation Vegetation Index (NDVI ^a); Length of Growing Period (LGP).
Geographic	Longitude (LONG), Latitude (LAT)
Topographic	Elevation (ALT)
Temperature	Air Temperature (AT ^a); Land Surface Temperature (LST ^a); Middle Infrared Reflectance (MIR ^a);
Water and moisture	Vapour Pressure Deficit (VPD ^a); Distance to Rivers (DRIV); Potential Evapo-transpiration (PEV)
Zoological	Cattle density (CAD); Holding density (HOD); Proportion of dairy cattle (PDC); Herd size (HES); Badger record density (BAD); Distance to nearest recorded badger presence (DBR).
Cattle movement	Total movements in (TM); Total movements in from infected areas (TMI); Proportion of movements from infected areas (PMI).
BTB persistence and local spread	Number of years of past BTB infection in the 5km cell (NYBTB); number of infected cells in the previous year in a 5 km radius doughnut window (DNT); Distance to the nearest cell with BTB present (DBTB)

a those variables were subjected to Fourier processing, and a series of additional data layers were produced (abbreviated suffix) : Mean (MN), Amplitude1 (AMP1), Amplitude2 (AMP2), Amplitude3 (AMP3), Phase1 (PH1), Phase2 (PH2), Phase3 (PH3), Variance of Mean (VM), Variance 1 (VAR1; variance 1 refers to the % of variance in the total signal accounted for by Fourier Component 1), Variance 2 (VAR2), Variance 3 (VAR3), Variance All (VARA), Minimum (MIN), Maximum (MAX), Range (RNG).

1.3 Case Selection

Logistic regression is very sensitive to the relative proportions of presence and absence values in the training data. With comparatively few records of BTB presence, especially in earlier years, the absence data tend to swamp the presence category and the logistic regression technique becomes unusable. Conventional analyses use case control techniques to select an appropriate sample of absence cases, whereby presence cases are matched by a variety of criteria, such as herd size, or holding type, to absence cases that are at the threshold distance, as determined by spatial autocorrelation analyses. This process was not feasible with the BTB data available, as there was insufficient information to categorise holdings effectively. As a result, sufficient absence locations to match the number of presence cases were selected from those areas with no disease by choosing cases at regular intervals within the database (such as every 20th case) when sorted by location (x and y) coordinates. This ensured an even geographical spread of absence cases, and avoided any selection bias in terms of any of the predictor criteria.

1.4 Multiple logistic regression model

The aim of these logistic regression models was to compare assess the effect of animal movement data over a previously published multiple logistic regression model. The analysis procedure was kept similar as in the previous study, and data were subjected to step-wise forward logistic regression analysis to establish the relationship between the predictor variables, including animal movement variables, and the presence or absence of disease.

Presence or absence of BTB was established using a threshold probability of 0.5 and the model predictive power was quantified by the % of correct presence, % of correct absence, overall % correct, and the kappa index of agreement¹², which ranges from 0 (no predictive skill) to 1 (perfect prediction), with values >0.4 regarded as acceptable and >0.75 as excellent¹³. At a resolution of 1 km, most positive squares are new disease because only a minor fraction of the detected cases are observed in squares where BTB was detected in the previous year. For example, out of the 443 squares where BTB was detected in 1995, only 42 occurred in a square with BTB reported in 1994; out of the 1054 squares with BTB detected in 2000, only 165 occurred in squares with report of BTB in 1999. Therefore the provided % of correct presence is quantitatively close to the % of squares with new disease correctly classified.

The contribution of each variable to the model was quantified by their Wald's statistic and only the 5 predictors with the highest Wald's statistic were shown in Table 1 of the main text, and the complete results are shown below in Supplementary Table 2.

These models were applied to the full 1 km resolution imagery to produce output maps predicting the probability of BTB presence throughout Great Britain (Fig. 2 of the manuscript)

Supplementary Table 2 Multiple logistic regression summary statistics and all predictors for models of BTB in Great Britain in 2002 and 2003 using cattle movement data from the current or two previous year, respectively Figures in brackets are Wald statistics

Year	2002	2003	2002	2003
Movement from	2002	2003	2000	2001
Var. avail. to model	100	100	100	100
Var. included	18	21	20	20
-2 Log likelihood	2671.8	3637.5	3000.3	3657.3
Chi-square	4444.4	2730.4	4174.2	2748.7
p	< 0.001	< 0.001	< 0.001	< 0.001
Correct pres. %	90.04	87.07	91.28	85.51
Correct abs. %	91.82	81.67	91.07	83.15
Overall correct %	90.94	84.39	91.17	84.33
Overall kappa	0.82	0.69	0.82	0.69
Variable 1	PMI (1033.42)	PMI (990.43)	PMI (842.13)	PMI (969.96)
Variable 2	DBTD00 (48.44)	ATVM (74.03)	DBTD00 (61.50)	ATVM (73.67)
Variable 3	ATVARA (47.17)	ATMN (44.55)	NDVIPH1 (34.63)	ATMN (44.89)
Variable 4	NDVIMAX (29.05)	ATPH1 (30.51)	DBR (33.97)	NDVIMN (33.59)
Variable 5	ATRNG (16.88)	MIRMN (28.18)	NDVIMAX (29.13)	VPDPH2 (25.85)
Variable 6	NDVIPH1 (15.00)	ATPH3 (22.62)	TMI (16.67)	DBTD01 (23.03)
Variable 7	PWO (13.91)	ATVAR2 (22.10)	HES (14.39)	LGP (19.73)
Variable 8	VPDMN (12.13)	VPDAMP3 (21.90)	ATRNG (13.63)	VPDPH3 (19.53)
Variable 9	HES (11.47)	NDVIMN (20.32)	PWO (11.61)	ATVAR2 (19.20)
Variable 10	VPDAMP2 (11.41)	DBTD01 (20.20)	ALT (11.56)	ATPH1 (18.22)
Variable 11	LSTAMP3 (9.99)	VPDPH3 (19.04)	PDC (11.11)	VPDAMP3 (11.92)
Variable 12	DBR (9.39)	LGP (16.76)	LSTMIN (9.84)	MIRMN (11.00)
Variable 13	MIRVAR3 (6.73)	VPDPH2 (11.83)	ATAMP1 (8.14)	POW (10.91)
Variable 14	ATVM (6.70)	VPDMN (10.69)	TMI (7.56)	TMI (9.78)
Variable 15	BAD (6.60)	DROADS (9.44)	ATVAR1 (7.15)	ATPH3 (8.63)
Variable 16	VPDPH2 (6.34)	VPDVAR1 (7.17)	MIRPH3 (6.31)	HES (6.98)
Variable 17	HPOP (4.17)	BAD (6.21)	HOD (6.04)	MIRMAX (6.96)
Variable 18	PDC (3.91)	POW (5.24)	VPDAMP2 (4.91)	VPDMN (6.67)
Variable 19	-	HES (4.62)	HPOP (3.99)	LSTPH1 (6.20)

Variable 20	-	PCU (4.55)	NDVIAMP2 (3.93)	PUR (6.16)
Variable 21	-	HPOP (4.55)	-	BAD (5.47)
Variable 22	-	-	-	TMI (5.23)

II Simulation models

The previous approach had two limitations in term of predicting BTB spread. Firstly, some of the predictors such as climate data are strongly correlated with each others, and different climatic predictors are obtained when the analysis is run for different years, so it is difficult to identify those predictors having the most stable relationship with BTB presence in time. Secondly, this method does not allow predictions more than one year ahead and it appeared useful to be able to generate predictions on the likely disease spread over longer period. The development of the simulation model involved the following steps.

II.1 Multi-annual logistic model

The aim of this statistical analysis was to use model BTB presence over a time series of BTB distribution such as to identify statistical associations relevant to the whole time series. The obtained variables and their parameters would then be used to simulate the spread over a time series starting from a known distribution. A similar statistical approach as described in the previous section was used with the difference that the dependent variable data set was here made of pooled observations from several years. In addition, this analysis required several adjustments detailed below.

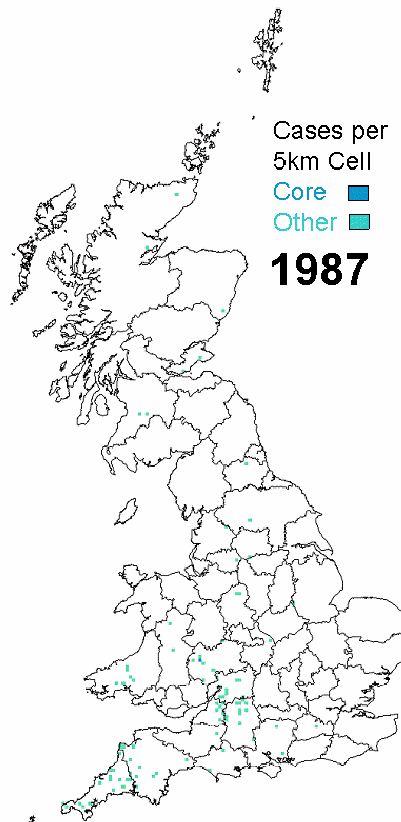
Additional variables. Some variables were added to account for the local persistence (Number of years of past BTB infection in the 5km cell; NYBTB), and for short-distance spread from neighbouring cells (number of infected cells in the previous year in a 5 km radius doughnut window; DNT);

Core and remote areas. The analysis was carried out separately for core areas (defined as 5 km squares where the disease has been present in 2 out of the three last years), and remote areas under the assumption that processes leading to disease occurrence might not be relatively equivalent in areas of endemicity, and in areas where the disease was rarely or never reported in the past. For example, movement of infective animals in a square have different implication in the likelihood of disease present in the next year if occurring in a square where the disease is repetitively observed (suggesting that BTB is locally persistent), than in a square where it was never reported before.

The 5 km spatial resolution of core areas resulted from an exploratory analysis of the spatial patterns in 1 km presence/absence data, showing spatial autocorrelation in disease status up to a distance of approximately 5 km. This result suggested that 5 km would be an adequate scale to define areas of persistence, i.e. in areas of endemicity. BTB would not necessarily be repetitively observed in the same 1 km cells, but much more likely in the same 5 km cells. A coarser resolution such as 10 km would have resulted in filtering out too much spatial variability.

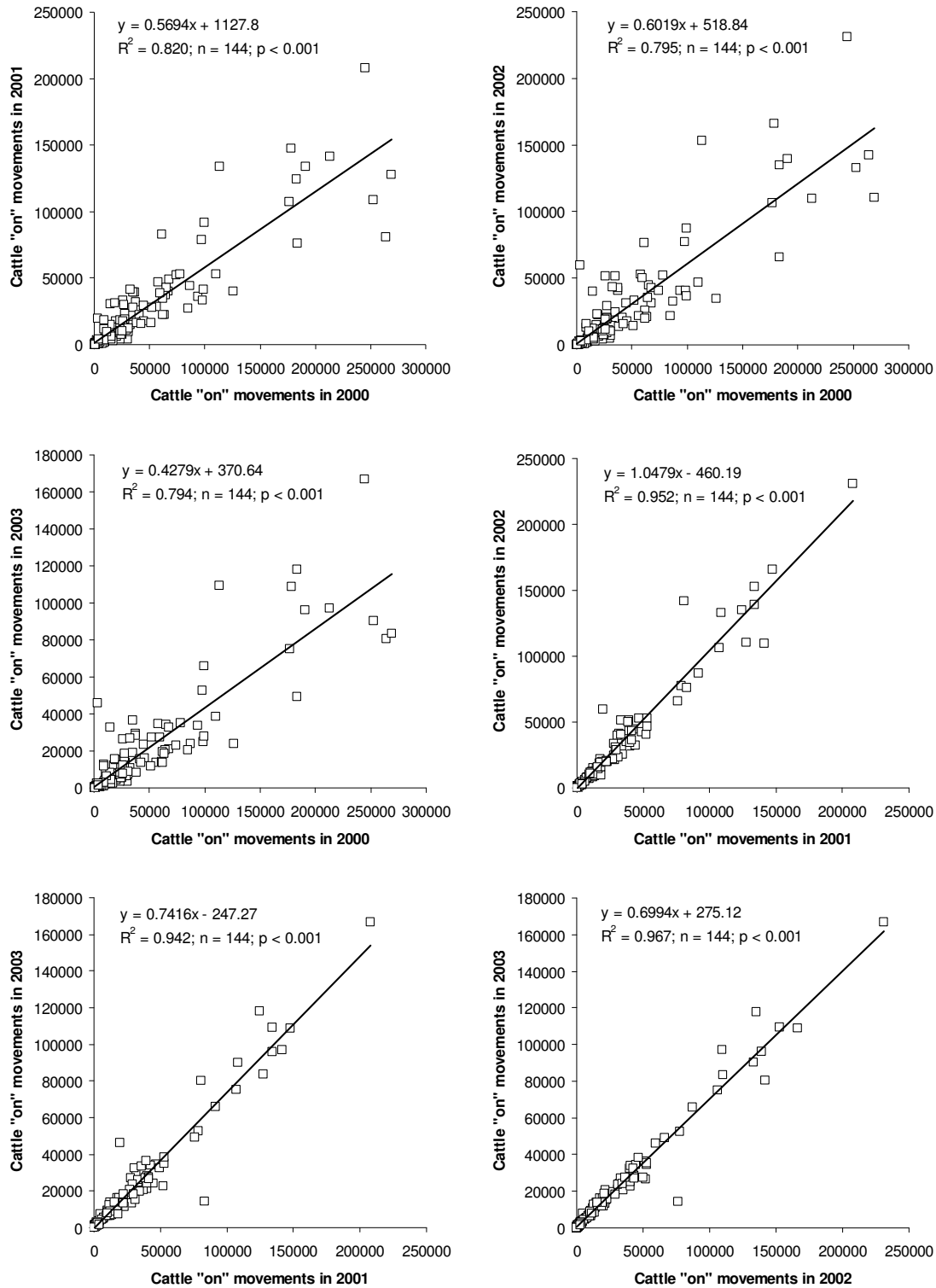
Core areas were thus defined as 5 km cells where BTB was detected in 2 out of three previous years, a definition resulting from exploratory analysis. We explored different definitions of core areas, and looked at the proportion of these squares defined as core areas where BTB was reported in the following year. For example, in 1999, BTB was detected in 311 squares out of the 652 squares (47.7 %) where TB was reported present at least one time in the period 1996-1998. In the same year, BTB was reported in 192 out of the 271 squares (70.8 %) where BTB was present in 2 out of the 3 previous years. Similarly, BTB was reported in 1999 in 73 out of the 93 squares (78.5%) where it had been reported every three previous years. These proportions, calculated for the year 2003 as a function of the period 2000-2002 are 53.3%, 73.0% and 79.5%, respectively. These figures show that the probability of having BTB detected in a given 5 km cell rises from around 50% to > 70% if one considers those squares where BTB was detected in more than one previous year, and does not increase so much if only cells where BTB was reported in all three previous years are considered. The definition of core areas, therefore, as 5 km cells where BTB was detected in 2 out of the 3 previous years, appeared to successfully distinguish between areas where the disease is reported occasionally from those where it is more persistently reported. These core areas changed every year as a function of the previous years observed distributions, as shown in Supplementary video 4.

Spread of BTB "core" areas in which the disease had been present for at least two of previous three years.



Supplementary Video 4. Animation of core area distribution 1985-2003 (open the supplementary video 4 to display the animation).

Animal movement data. The multi-annual model required the estimate of inward movement in years when animal movement data were not yet collected and compiled. We made the assumption that annual figures of annual movement were stable in time, a that a fixed movement index – the proportion of total movement between 2000 and 2003 from infected areas - could be incorporated into the multi-annual predictor suite. This assumption was based on the comparison between 2000, 2001, 2002 and 2003 movement data, and can be visually assessed in Supplementary Figure 2 and Supplementary videos 2 & 3 (data section). This assumption was quantified by the tight regression between yearly movement data aggregated by county (Supplementary Figure 2), and a good fit was also obtained when comparing movement data at the 5 km cell resolution (R^2 equals to 0.53, 0.49, 0.46, 0.80, 0.91 and 0.92 for the pairs of years 2000-2001, 2000-2002, 2000-2003 2001-2003, 2001-2002 and 2002-2003 respectively).



Supplementary Figure 2. Regression between county-level total inward animal movements for the years 2000, 2001, 2002 and 2003.

II.2 Variable selections.

As previously mentioned, the aim of the multi annual model was to identify stable statistical associations relevant to the whole time series. Given that the main scope of the model is to generate predictions, any variable that substantially improves the predictive power of the model can be incorporated. However, variables with a clear and interpretable link with the process of interest are to be preferred because these are more likely to present similar association in the future, whereas some other predictors may present a strong but fortuitous statistical association in the studied data set time series, which may not be such an adequate predictor for future distributions. Therefore, we first tested the variables relating to local persistence (Number of years of past BTB infection in the 5km cell, NYBTB), to short-distance spread (number of infected cells in the previous year in a 5 km radius doughnut window DNT), to animal movements (Total movements in TM, Total movements in from infected areas TMI; Proportion of movements from infected areas PMI), and to cattle density (CAD). Then other variables were entered into the model using a standard forward entry stepwise procedure, until no significant predictors could be entered into the model. This first selection procedure resulted in 13 and 31 variables included in the core area, and remote areas multiple logistic regression models, respectively,

We aimed to restrict the model to variables with the highest predictive power, and the change in log likelihood upon removal of each variable was estimated. The second step involved removing variables which removal resulted in a change in model log likelihood lower than 1%. At the end of this step, the core area model included the following variables (% change in model log likelihood upon removal in brackets):

NYBTB: Number of years of past BTB infection in the 5km cell (17.96%);

DNT: Number of infected cells in the previous year in a 5 km radius doughnut window (5.41%);

whereas the remote areas model included the 10 following variables:

DNT: Number of infected cells in the previous year in a 5 km radius doughnut window (16.48%)

PMI: Proportion of movements from infected areas (13.18%)

NYBTB: Number of years of past BTB infection in the 5km cell (7.09%)

NDVIMN: Normalised Deviation Vegetation Index Mean (6.97%)

NDVIMAX: Normalised Deviation Vegetation Index Maximum (4.21%)

CAD: Cattle density (2.63%)

NDVIRNG: Normalised Deviation Vegetation Index Range(1.33%)

PCU: Percentage cultivation and managed grassland (1.29%)

In the last step, we aimed to reduce the risk of including variables irrelevant to future BTB distributions by verifying that variables included in the model resulted from associations that could be considered as stable in time, i.e. significant for each year. This verification was necessary because some variables fortuitously associated with one or a few single year BTB distribution may well appear significant in the multi-annual model if their association is strong enough, whereas the association with other years would be much weaker. The variables selected at the previous steps were tested in annual multiple logistic regression models within each year, and variables not significant at the p level of 0.05 or more than one year were excluded from the model. The significance level of each variable each year for core and remote areas is presented in Supplementary Table 3 below:

Supplementary Table 3 Significance of the variables in the core and remote areas annual multiple logistic regressions of BTB presence from 1997 to 2003.

Year	1997	1998	1999	2000	2001	2002	2003
Core areas							
NYBTB	0.071	<0.001	0.008	<0.001	0.002	<0.001	<0.001
DNT	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Remote areas							
NYBTB	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

DNT	<0.001	<0.001	<0.001	<0.001	0.002	<0.001	<0.001
PMI	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
NDVIMN	0.003	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
NDVIMAX	0.198	0.320	0.077	0.076	<0.001	0.013	0.072
CAD	<0.001	<0.001	<0.001	0.075	0.001	<0.001	<0.001
NDVIRNG	0.076	0.587	0.571	0.004	<0.001	<0.001	<0.001
PCU	0.003	0.008	0.530	<0.001	<0.001	<0.001	<0.001

The final multi annual model therefore included the variables NYBTB and DNT in the core areas, and the variables NYBTB, DNT, PMI, NDVIMN, CAD and PCU in remote areas (see variables details above). Details of this model are found in Table 2 of the text.

11.3 Simulations.

Simulation algorithm. First, infection probability of each cell was estimated as a logistic function of a series of predictors identified using the multi annual multiple logistic regression models detailed in the previous sections. Second, a layer of random numbers (uniform distribution) was generated and cells with a random number lower than their infection probability were set as BTB-present. Third, each cell's BTB status was updated and the algorithm re-iterated to simulate the spread in the next year.

The algorithm started with the observed distribution of BTB in 1997 and iterated until the target year (2003, 2004 or 2005). This set of n hypothetical distributions of BTB in n consecutive years constitutes a single run of which 500 were carried out, and the number of BTB-presence within each 1 km square was averaged over the 500 runs. These were compared to the observed distribution at the same scale, by estimating the model log-likelihood and McFadden's pseudo-R² statistic as in a standard logistic regression. The simulation model was developed in R¹⁴.

Movement kernel model. Processing constraints prevented the proportion of inward movement from infected areas from being generated on the fly during the simulation process, and it was thus necessary to identify a surrogate variable to animal movement from infected area.

The animal movement kernel, averaged over the years 2000-2003, was modelled as a function of distance. The obtained function could then be used to transform geographical distance to the nearest square infected in the previous year into an surrogate index of potential inward movements. The movement data were grouped in 160 distance classes, each representing the frequency of movements for each 5 km intervals going from 0-5 km (first class) to 795-800 km (last class). The movement frequency distribution was considered as a multinomial distribution, and several reduced parameter multinomial models were tested in their goodness of fit by estimating their likelihood ratio statistic as

$$G^2 = 2 \sum_{j=1}^k y_j \ln \left(\frac{y_j}{\hat{E}_j} \right)$$

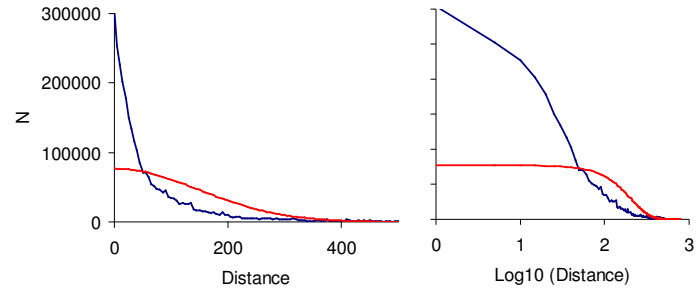
where y_j is the observed frequency of class j , \hat{E}_j is the ML-estimated expected frequency of the class j , and k is the number of classes. The different parametric forms which were tested are detailed in Supplementary Figure 3 below.

2 parameters

Model A

$$\hat{E} = n.\exp(-3.626 - 2.29 \cdot 10^{-5} D^2)$$

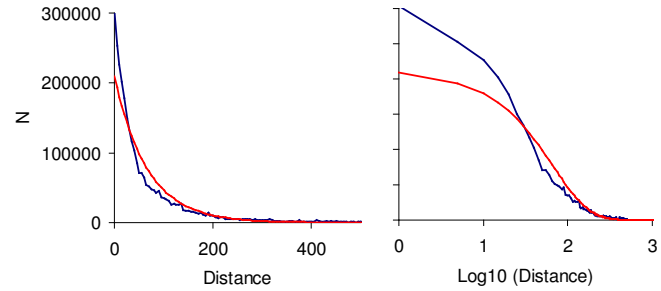
$$G^2 = 1\,996\,636$$



Model B

$$\hat{E} = n.\exp(-2.655 - 0.0137 D)$$

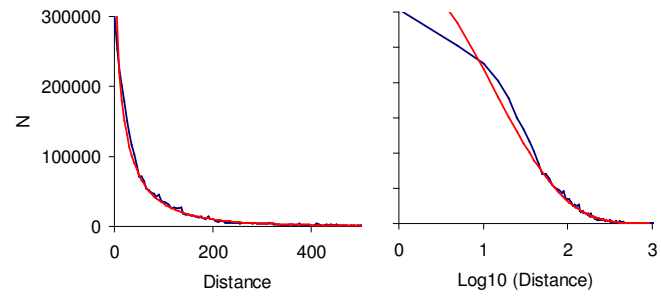
$$G^2 = 275\,398$$



Model C

$$\hat{E} = n.\exp(-1.711 - 0.277 D^{0.5})$$

$$G^2 = 88\,100$$

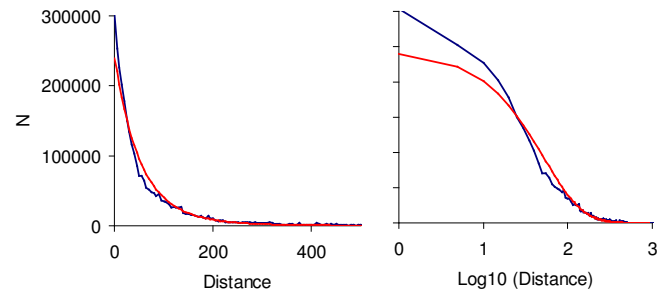


3 parameters

Model D

$$\hat{E} = n.\exp(-2.442 - 0.0200 D + 1.637 \cdot 10^{-5} D^2)$$

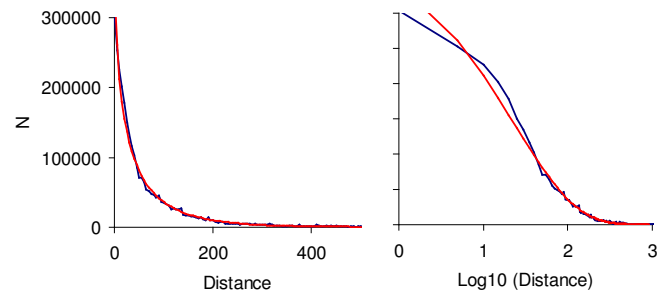
$$G^2 = 113\,846$$



Model E

$$\hat{E} = n.\exp(-1.920 - 0.00366 D - 0.210 D^{0.5})$$

$$G^2 = 44\,283$$

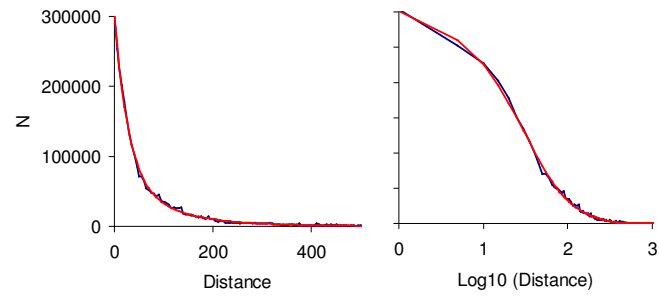


4 parameters

Model F

$$\hat{E} = n.(\exp(-2.498 - 0.0344 D) + \exp(-3.853 - 0.00909 D))$$

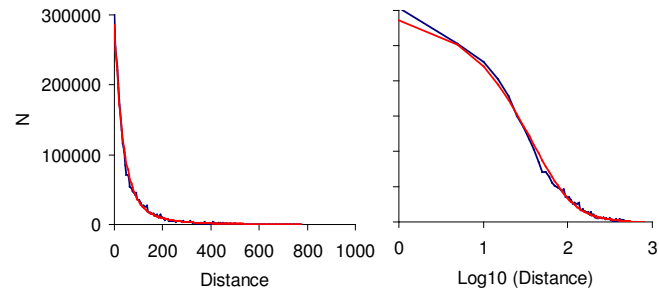
$$G^2 = 29\,075$$



Model G

$$\hat{E} = n.\exp(-2.313 - 0.0263 D - 5.528 \cdot 10^{-5} D^2 - 5.118 \cdot 10^{-8} D^3)$$

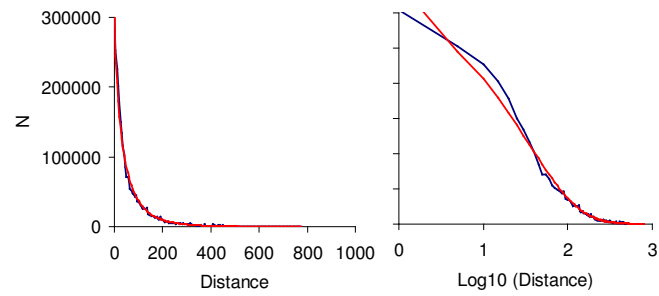
$$G^2 = 33\,511$$



Model H

$$\hat{E} = n.\exp(-2.148 - 0.0100 D - 0.129 D^{0.5} - 6.884 \cdot 10^{-6} D^2)$$

$$G^2 = 53\,845$$

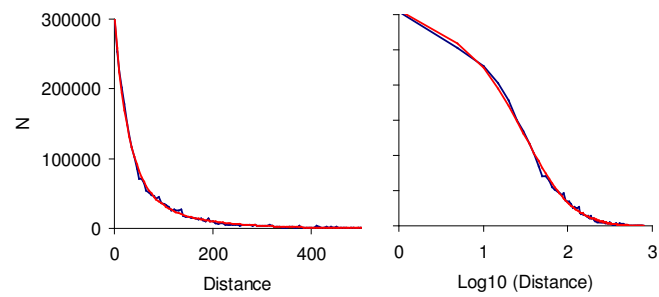


5 parameters

Model I

$$\hat{E} = n.((\exp(-2.629 - 0.0347 D) + \exp(-3.381 - 0.00742 D - 0.0563 D^{0.5})))$$

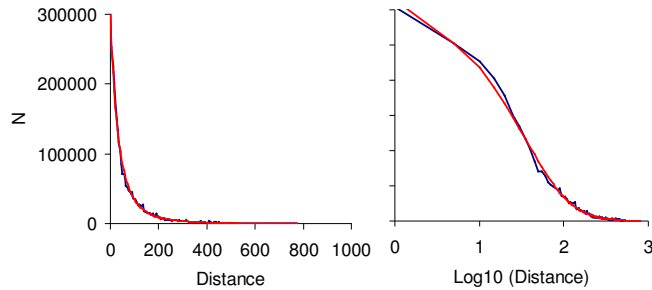
$$G^2 = 28\,702$$



Model J

$$\hat{E} = n \cdot \exp(-2.221 - 0.0210 D - 0.0504 D^{0.5} - 4.290 \cdot 10^{-5} D^2 - 4.002 \cdot 10^{-8} D^3)$$

$$G^2 = 29\,138$$



Supplementary Figure 3. Movement kernel models, D is the distance expressed in meters, and G^2 is the model log-likelihood ratio.

From the different parametric forms detailed in the Supplementary Figure 3, the first model with 5 parameters (Model I) provided the best adjustment to the observed movement kernel: the incorporation of a 5th additional parameter resulted in a statistically better model (difference in LL ratios with Model F is 373, sig. of the change < 0.001) and resulted in a reduction of 1.28% of the LL ratio as compared to the best 4-parameters model. This model (model I) was used at simulation time to transform the distance to the nearest 1 km square infected in the previous year, and this variable was forced into the multi annual multiple logistic regression model in place of the movement variable (PMI) in the model presented in Table 2 of the text.

Parameters. The parameters of each variable used in the simulations were estimated by a multi annual logistic regression model estimated over a given period, with an **unbalanced dataset** (full 1 km resolution imagery). These regressions included the year as complementary variable to account for the increase in cases as a function of time. The model parameters for the different periods, and for the different models (core/remote areas, model with PMI or TDBTB) are presented in the Supplementary Table 4 below.

Supplementary Table 4 Parameters used in the simulation models for the predictors identified in the multi-annual logistic regression analysis.

Period for parameters estimation: 1997-2002

Core area		Remote area (with PMI)		Remote area (with TDBTB)	
NYBTB	0.0723	NYBTB	0.329	NYBTB	0.241
DNT	0.161	DNT	0.365	DNT	0.116
YEAR	0.0770	PMI	4.97	TDBTB	56.6
CONSTANT	-157	NDVIMN	0.00651	NDVIMN	0.00277
-	-	CAD	0.00605	CAD	0.00185
-	-	PCU	0.0105	PCU	0.00752
-	-	YEAR	0.134	YEAR	0.163
-	-	CONSTANT	-286	CONSTANT	-340

Period for parameters estimation: 1997-2001

Core area		Remote area (with PMI)		Remote area (with TDBTB)	
NYBTB	0.0620	NYBTB	0.298	NYBTB	0.230
DNT	0.224	DNT	0.398	DNT	0.115
YEAR	-0.122	PMI	5.58	TDBTB	72.9
CONSTANT	240	NDVIMN	0.00767	NDVIMN	0.00362
-	-	CAD	0.00625	CAD	0.00116
-	-	PCU	0.00782	PCU	0.00555
-	-	YEAR	-0.122	YEAR	-0.0673
-	-	CONSTANT	224	CONSTANT	116

Period for parameters estimation: 1997-2000

Core area		Remote area (with PMI)		Remote area (with TDBTB)	
NYBTB	0.0714	NYBTB	0.284	NYBTB	0.233
DNT	0.230	DNT	0.408	DNT	0.138
YEAR	0.0818	PMI	6.03	TDBTB	64.8
CONSTANT	-166	NDVIMN	0.00825	NDVIMN	0.00412
-	-	CAD	0.00575	CAD	0.00102
-	-	PCU	0.00669	PCU	0.00475
-	-	YEAR	0.0594	YEAR	0.100
-	-	CONSTANT	-138	CONSTANT	-218

III Discussion

Inclusion of herd density variables in the model.

Herd size and proportion of dairy cattle were available to the authors for 1995 and 1997, as was holding density, which was used as a surrogate for herd density. The earlier work showed that none of these predictors were found within the top 5 predictors of BTB density, and only one within the first 10 for one of the years (1995). Nevertheless, they were included in the 2002 and 2003 models and whilst they were indeed found to be significant they had very weak predictive power (Table 2). Inclusion of these variables did not materially improve the models (as indicated by the % correct predictions), and indeed in some cases slightly reduced the goodness of fit. Multi-annual models incorporating the available herd variables also showed that, while significant, their predictive value was even lower (expressed in terms of change of model log likelihood upon removal). We therefore concluded that these variables are not essential components of the multi-annual models constructed here. This does not, however, preclude their potential importance in a biological sense, as their impact may well be more evident within local environments than the rather large geographical ranges covered by the present models.

Uncertainty resulting from the testing regime or imperfect BTB skin test.

Where herds are tested less frequently, especially where testing is restricted to once every three or four years, Defra recognises there is a risk that some geographical areas with cattle may not get any testing for that period, implying that introduced infection could be festering for several years before detection. To prevent this happening two measures are in place:

Although not hugely sensitive as a method of detection, all cattle carcasses for human consumption and those in the over 30 month scheme will be subject to a post mortem examination designed to pick up lesions visible to the eye. 20%+ of confirmed cases are detected in this way in the 3-4 year testing areas, compared to about 10% of cases in yearly tested areas.

Secondly, DVMs are instructed to make sure that routine testing is not concentrated in any one geographical area to the detriment of any other area in any one year, i.e. testing is spread throughout an area over time. Therefore, as TB tends to cluster, it is unlikely that all members of a cluster will remain undisclosed for any length of time.

This does not invalidate our results. Variation in testing frequencies may affect annual BTB presence/absence patterns, but given a time series of distribution, the main effect should be a level of uncertainty on the dating of the detected case in those low-frequency testing areas, rather than an effect on their absolute number because undetected cases missed in any given year are presumably compensated by new detected cases which were missed in the previous year(s). The consequence of uncertainty in the dating of detected cases in some areas on our result would be to reduce the strength of the association we find with movement data (which are correctly dated). In other words, variations in testing regime add noise to the temporal resolution of case data in some areas, and one could actually expect a stronger association with movement data if testing was homogeneously annual. In addition, the robustness of the association with movement data with regards to testing regime is further supported by the multi-annual models that includes several years of testing data, and where any undetected case would appear, and be accounted for, later in the analysed time-series.

The imperfect sensitivity of the skin test may have some effect, but we have no reason to believe that this effect is not spatially homogeneously distributed, and therefore simply adds to the unexplained variability.

Incorporation or exclusion of cells without Cattle

The data set used to build the regression and simulation models included all cells, and we had several reasons for not excluding cells on the basis of the available data on cattle distribution. First, the testing regime resulted in a dataset containing some uncertainty on the BTB status, which is not homogeneously distributed in time and space and cannot guarantee to cover 100% of the cattle in Great Britain. The available distribution data are either as densities per administrative area (Scotland and Wales) or holding level census data for England. The former will tend to have rather few zero values as there will be few administrative areas with no cattle at all. Even the English holding data are not fully comprehensive as they exclude the 'minor' holdings. Thus apparent zeros may not be real. The TB testing data are also an unreliable source of zeros, as not every holding was tested each year. Given these uncertainties, it was preferred to build models making the least assumptions on the distribution of negatives. Second, this option facilitates implementing temporal models as it does not rely on the availability of annual cattle distributions, which is also subject to change in the future. Third, the fact that there was a very low proportion of negatives falling in areas with an estimated cattle density of zero (< 5 %) indicated that excluding those points (provided they are true zero density areas) has little impact on the predictions, or on the main results of the paper, i.e. movement data can be used to predict the spread of BTB.

This has been confirmed by repeating the regression analyses (annual and multi-annual) with the exclusion of zero cattle density BTB points, and the results are presented below in Supplementary Table 5 and Supplementary Table 6, which compare to Table 1 and Table 2 of the main text. As can be seen from the comparison, the model excluding the points with zero cattle density did not result in model improvement as measured by the overall % correct, or correct pres. %, as the model with the zero points included had a slightly better predictive power. The same observation can be made from the comparison of Table 2 and Table 2b regarding the multi-annual models, with both Tables showing very similar results, and slightly better predictions with the models incorporating all the sample points including those with zero cattle density.

In summary, the authors suspect that the zero data are not sufficiently reliable to justify excluding them from the analysis, especially in the light of the very minor (and indeed slightly negative) impact on the results of doing so.

Supplementary Table 5 Multiple logistic regression summary statistics and first five predictors for models of BTB in Great Britain in 2002 and 2003 using cattle movement data from the current or two previous year, respectively (excluding zero cattle cells).

Year	2002	2003	2002	2003
Movement from	2002	2003	2000	2001
Var. avail. to model	100	100	100	100
Var. incl. in the model	18	22	19	23
-2 Log likelihood	2673.3	3605.4	2973.2	3626.7
Chi-square	4331.0	2674.4	4074.3	2679.5
p	< 0.001	< 0.001	< 0.001	< 0.001
Correct pres. %	90.43	86.54	90.83	85.04
Correct abs. %	91.39	81.25	91.25	82.82
Overall correct %	90.92	83.91	91.03	83.93
Overall kappa	0.82	0.68	0.82	0.68
Variable 1	PMI ^a (1048.2)	PMI ^a (990.7)	PMI ^a (856.6)	PMI ^a (956.4)
Variable 2	DBTD00 ^b (55.5)	ATPH1 ^f (46.6)	DBTD00 ^b (47.8)	NDVIMN ^g (38.9)
Variable 3	HES ^c (28.7)	NDVIMN ^g (36.8)	NDVIPH1 ⁱ (38.3)	ATVM ^l (25.3)
Variable 4	NDVIPH1 ^d (21.2)	ATAMP1 ^h (35.2)	NDVIMAX ^j (33.9)	VPDPH2 ^m (21.6)
Variable 5	ATRNG ^e (20.4)	DBTB01 ^b (30.3)	DBR ^k (33.3)	ATVAR2 ⁿ (20.2)

a Proportion of movements from infected areas; b Distance to the nearest cell with BTB present in year YY; c Herd size; d Normalised Deviation Vegetation Index (phase 1); e Air temperature (range); f Air temperature (phase 1); g Normalised Deviation Vegetation Index (mean); h Air temperature (amplitude 1); i Normalised Deviation Vegetation Index (phase 1); j Normalised Deviation Vegetation Index (max); k Distance to nearest recorded badger presence; l Air temperature (variance of mean); m Vapour pressure deficit (phase 2); n Air temperature (var 2)

Supplementary Table 6 Multi-annual multiple logistic regression of BTB occurrence in Great Britain from 1990 to 2003 (excluding zero cattle cells)

	Core	Remote with movement data	Remote with transf. dist
Var. avail. to model	100	100	100
Var. incl. in the model	2	6	6
-2 Log-likelihood	10261.1	7091.4	6186
Chi-square	1276.9	5437.6	6343
p	< 0.001	< 0.001	< 0.001
Correct pres. %	56.26	78.81	87.44
Correct abs. %	77.96	86.47	84.58
Overall correct %	67.08	82.75	85.97
Overall Kappa	0.34	0.65	0.72
Variable 1	NYBTB ^a (660.9)	NYBTB ^a (129.1)	NYBTB ^a (79.7)
Variable 2	DNT ^b (270.3)	DNT ^b (310.2)	DNT ^b (184.2)
Variable 3	-	PMI ^c (437.7)	TDBTD ^g (1117.1)
Variable 4	-	NDVIMN ^d (81.4)	NDVIMN ^d (48.1)
Variable 5	-	CAD ^e (167.4)	CAD ^e (43.6)
Variable 6	-	PCU ^f (120.9)	PCU ^f (65.4)

a Number of years of past BTB infection in the 5km cell; b number of infected 1 km cells in the previous year in a 5 km radius doughnut window; c Proportion of movements from infected areas; d Normalised Deviation Vegetation Index Mean; e Cattle density; f Percentage cultivation and managed grassland; g Transformed distance to the nearest 1 km square with BTB reported in the previous year. Figures in brackets are Wald statistics.

Proportion of movement from infected areas is a better predictor than absolute numbers.

It is believed that this may possibly relate to two non-exclusive effects.

Firstly, this result might relate to a effect of dilution. Consider two squares, each receiving an absolute number of 10 animals from infected areas, out of 50 and 500 animals entering the square respectively. If the cattle population in the square is 100 animals, the first case will result in 10/160 animals from infected areas, whereas the second case will result in 10/610 animals. The chances of contact between potentially infective and susceptible animals are clearly higher in the first case than in the second where the animals that may possibly transmit BTB are somehow “diluted” among the susceptible animals. This effect may only be significant if the number of animals in an area is relatively low as compared to the movement figures.

A second effect may also apply when local population is high as compared to the movement of animals. The total number of inward animal movement is an indicator of local inward and outward animal flows, because areas with high inward movements generally correspond to areas of high outward movement also, otherwise the number of animals in a specific area would continually increase over time. The higher significance of PMI therefore implies that a given number of animal movements from infected areas pose a higher risk of transmission if they occur in an areas with low inward and outward cattle movements (i.e. low level of trade), as compared to areas with high flows and trade of animals, where the infective animals are more likely to be moved away within a short period. In other words, those animals moving in from infected areas may only have a transient stay in those high flow and trade areas. By the same token, it is likely that low levels of on movements are matched by low levels of off movements and stays tend to be longer, and with it the chances of transmitting BTB. The

critical epidemiological variable would therefore be the number of potentially infective animals multiplied by their length of stay.

This can also be expressed in probabilistic terms using a binomial approach. Consider an area with a high local population relative to the number of inward movements, say 10,000 cattle, with 50 animals moving in from infected areas, out of 200 animals in the first case (low inward flow), and out of 2000 animals in the second case (high inward flow). For the purposes of illustration, it is assumed that inward movements are equal to outward movements (200 and 2000 respectively). In the first case, after inward movements, we have 50 out of 10,200 animals, thus a proportion of 0.4902 % of potentially infective animals. Take randomly 200 of these 10,200 animals, and a binomial model will predict a probability of 0.374 chance to export none of these potentially infective animals, i.e. that those animals that came from infected areas would all remain there. In the second example, the proportion of animals from infected areas after inward movements locally is equal to 0.417 % (50/12,000), but taking randomly 2000 of these cattle, and the binomial model predicts only a probability of 0.000236 to export none of these potentially infective cattle. The probability that animals moving in from infected areas stay locally is thus much higher in the first example involving a low flow of absolute inward and outward movements.

The PMI variable may thus account for the fact that an equivalent number of animals moving in from infected areas pose a higher epidemiological risk if they are part of small flows of animals. Clearly, there is a need to disentangle these possible effects, and one possible approach for future work could be to track cattle movements as well as their length of stay, such as to separate animals transient in an area from those that remain for a longer period, and are thus at higher risk of transmitting BTB if they are themselves infected.

IV References

1. Wint, G. R., Robinson, T. R., Bourn, D. M., Durr, P. A., Hay, S. I., Randolph, S. E. & Rogers, D. J. Mapping bovine tuberculosis in Great Britain using environmental data. *Trends Microbiol.* **10**, 441-444 (2002).
2. CIS. *Countryside Information System CD-ROM, Version 6.0*. (Produced by WS Atkins and Dart Computing by the Department of the Environment, Transport and the Regions, 1999).
3. Eidenshink, J.C. & Faundeen, J.L. The 1km AVHRR global land data set – 1st stages in implementation. *Int. J. Remote Sensing* **15**, 3443–3462 (1994)
4. Teillet, P.M. *et al.* An evaluation of the global 1-km AVHRR land dataset. *Int. J. Remote Sensing* **21**, 1987–2021 (2000)
5. Hay, S. I. & Lennon, J. J. Deriving meteorological variables across Africa for the study and control of vector-borne disease: a comparison of remote sensing and spatial interpolation of climate. *Trop. Med. Int. Health* **4**, 58–71 (1999)
6. Price, J.C. Land surface temperature measurements from the split window channels of the NOAA 7 advanced very high resolution radiometer. *J. Geophys. Res.* **89**, 7231–7237 (1984)
7. Rogers, D.J. Satellites, space, time and the African trypanosomiases. *Adv. Parasitol.* **47**, 129–171 (2000)
8. Wilson, G. *et al.* *Changes in the British badger population, 1988 to 1997* (People's Trust for Endangered Species, London, 1997)
9. Arnold, H. R. *Atlas of Mammals in Britain* (HMSO, London, 1993)
10. Wint, G. R. W., Gilbert, M., Bourn, D. M., Mitchell, A. & Clifton-Hadley, R. *Exploratory Investigation of Cattle Movement Records in Britain to Enhance Animal Disease Surveillance and Control Strategies* (Final Report to DEFRA on Research Project CSA6400/SE3034, http://www2.defra.gov.uk/research/project_data, 2004).
11. Mitchell, A., Bourn, D., Mawdsley, J., Wint, W., Clifton-Hadley, R. & Gilbert, M. Characteristics of cattle movements in Britain – an analysis of records from the Cattle Tracing Scheme. *J. Anim. Sci.* (In press)
12. Carstensen, L.W. A measure of similarity for cellular maps. *The American Cartographer* **14**, 345–358 (1987)
13. Landis, J. R. & Koch, G.C. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977)
14. R Development Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, Vienna, <http://www.R-project.org>, 2004)