Europ. J. Agronomy 74 (2016) 173–184

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

Mapping crop diseases using survey data: The case of bacterial wilt in bananas in the East African highlands

H. Bouwmeester[a,*], G.B.M. Heuvelink[b,c], J.J. Stoorvogel[c]

[a] GeoSpace, Roseboomlaan 38, 6717 ZB Ede, The Netherlands
[b] ISRIC - World Soil Information, PO Box 353, 6700 AJ Wageningen, The Netherlands
[c] Wageningen University, Soil Geography and Landscape Group, PO Box 47, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 13 July 2015
Received in revised form
21 December 2015
Accepted 22 December 2015

Keywords:
Agriculture
Geostatistics
Indicator regression kriging
Sampling design
Spatial aggregation

ABSTRACT

Globally, crop diseases result in significant losses in crop yields. To properly target interventions to control crop diseases, it is important to map diseases at a high resolution. However, many surveys of crop diseases pose challenges to mapping because available observations are only proxies of the actual disease, observations often are not normally distributed and because typically convenience sampling is applied, leading to spatially clustered observations and large areas without observations. This paper addresses these challenges by applying a geostatistical methodology for disease incidence mapping. The methodology is illustrated for the case of bacterial wilt of banana (BWB) caused by *Xanthomonas campestris pv. musacearum* in the East African highlands. In a survey using convenience sampling, 1350 banana producing farmers were asked to estimate the percentage yield loss due to bacterial wilt. To deal with the non-normal distribution of the data, the percentages were classified into two binary variables, indicating whether or not the disease occurred and whether or not the yield loss was severe. To improve the spatial prediction of disease incidence in areas with low sampling density, the target variables were correlated in a logistic regression to a range of environmental variables, for which maps were available. Subsequently, the residuals of the regression analysis were interpolated using simple kriging. Finally, the interpolated residuals were added to the regression predictions. This so-called indicator regression kriging approach resulted in continuous maps of disease incidence. Cross-validation showed that the method yields unbiased predictions and correctly assesses the prediction accuracy. The geostatistical mapping is also more accurate than conventional mapping, which uses the mean of observations within districts as the predicted value for all locations within the district, although the accuracy improvement is not very large. The maps were also spatially aggregated to district level to support regional decision-making. The analysis showed that the disease is widespread on banana farms throughout the study area and can locally reach severe levels.

1. Introduction

Despite years of agricultural research, diseases, pests, and weeds still seriously hamper agricultural productivity. In literature reviewed by Savary et al. (2012) direct yield losses caused by diseases, animals, and weeds are estimated to be responsible for losses ranging between 20% and 40% of global agricultural productivity. Large yield gaps, i.e., substantially lower actual yields compared to their potential (Van Ittersum et al., 2012; Spiertz, 2012), are particularly found in regions with low external inputs, such as Eastern Africa. Reducing the yield gap in Africa scores high on the political agenda, considering the serious problems around food security (Ozor et al., 2014) and the importance of agriculture for the national economies (agriculture employs the majority of the workforce and contributes significantly to the gross domestic product) (AfDB et al., 2015). Crop diseases are among the main causes of yield gaps (Van Ittersum et al., 2012; Wang et al., 2015). To allow extension officers and policy makers to act better informed and to target pests and diseases efficiently, it is necessary to map the distribution and extent of crop diseases. This supports the design and allocation of proper management measures and effective policies. Regional datasets on crop diseases collected through farm surveys provide valuable information to generate disease maps. However, these datasets have a number of characteristics that hamper mapping (Bouwmeester et al., 2012):

* Corresponding author.
E-mail address: h.bouwmeester@solum.nl (H. Bouwmeester).

http://dx.doi.org/10.1016/j.eja.2015.12.013
1161-0301/

- Surveys rarely focus on diseases directly and often make use of proxies based on symptoms or estimated yield loss. These proxies are subsequently interpreted as indicators of disease incidence. For example, Legg et al. (2006) reviewed the literature reporting on the spread of Cassava Mosaic Virus and show that a range of different methods to describe the virus incidence were used, ranging from ELISA-based diagnostics, nucleic acid-based techniques, but also the registration of whitefly, the vector of CMV, and disease symptoms. Brentu et al. (2012) used the percentage of affected citrus trees with Citrus Black Spot as a proxy for crop loss. Disease symptoms also do not distinguish between different causal agents and may be the result of multiple interacting microbes rather than a single one, which hampers attribution of disease incidence to a single cause, such as BWB.
- Data on crop diseases are often qualitative (e.g., low, medium, high disease severity) or binary (absent or present). For example, Zinga et al. (2013) scored the severity of Cassava Mosaic Disease from 1 (asymptomatic) to 5 (most severely diseased).
- If data are quantitative, they are rarely normally distributed and often present a large number of zeros. These data characteristics require specific statistical methodologies and tailored approaches.
- Data are mostly collected through a convenience sample (e.g., Thompson et al., 2011) as terrain accessibility is problematic and resources do not allow for sampling design optimization for mapping purposes. This often results in an uneven geographical distribution of points and/or clustered data.
- The data often present a high short-distance variability, as illustrated by a case study for palm oil by Rakib et al. (2014).

The above limitations of survey data are often the result of choices in the survey design while dealing with the trade-off between resource availability and data quality requirements. Limited resources may force surveyors to measure proxies of a disease and use convenience sampling. Nevertheless, despite the practical problems with data collection, several studies used survey data to map crop diseases in Eastern Africa. Tusheremereirwe et al. (2006) created maps of bacterial wilt in banana in Uganda, Night et al. (2011) created maps of cassava mosaic disease and cassava brown streak disease in Rwanda, and Legg et al. (2006) created maps for the same cassava diseases for East and Central Africa. Most studies calculate the arithmetic mean of the crop disease parameter in an area delineated by administrative boundaries. However, this conventional method does not address the limitations listed above, and does not quantify the uncertainty associated with the maps.

The objective of this study is to apply a geostatistical methodology to map disease patterns while dealing with the aforementioned limitations in survey design as adequately as possible. Similar approaches but with different limitations and applications are reported in Munar-Vivas et al. (2010), Park et al. (2012), Lamichhane et al. (2013), and Rees et al. (2014). The methodology is illustrated with a case study on bacterial wilt of banana (BWB) caused by *Xanthomonas campestris pv. musacearum* in the East African highlands. BWB is considered to be one of the major threats to banana yields in the region (Blomme et al., 2014; Biruma et al., 2007). The East African highlands is the largest banana producing and consuming region of Africa (FAO, 2012). Up to 60% of the caloric intake of the rural population consists of banana products and bananas are an important cash–crop (Abele et al., 2008). BWB affects all banana cultivars and causes rotting of the banana bunch, yellowing of the leaves and eventually dying of the mother plant. It causes immediate yield losses but also impacts the plant production cycle on the long term. BWB was introduced to Africa in the 1960s in Ethiopia, where it stayed and did not spread until it was reported in Uganda in 2001, from where it quickly spread to the Democratic Republic of Congo, Rwanda, Tanzania, Kenya, and Burundi (Tripathi et al., 2009). Blomme et al. (2013) identify regular surveillance, the supply of clean planting material and phytosanitary measures as crucial to control diseases such as BWB and maintain a healthy banana sector in Africa. Maps on disease distribution and damage are required to guide these management measures. We explore the use of indicator regression kriging for mapping and compare the results with conventional mapping methods. The performance of both mapping methods is evaluated using a visual interpretation and cross-validation.

2. Materials and methods

2.1. Study area

The study area (Fig. 1) is part of the East African highlands and includes Rwanda, Burundi and the main banana growing areas in Western Kenya, Southern Uganda and Northern Tanzania. The East African highlands are part of the African Rift System with Lake Victoria as the central basin. The area has a diverse agro-ecology, resulting from large variation in altitude (546–4610 m above sea level (m.a.s.l.); CGIAR-CSI, 2008), mean annual rainfall (492–2250 mm; Hijmans et al., 2005) and mean annual temperature (from 3.0 to 25.9 °C; Hijmans et al., 2005). A large part of the area is very suitable for growing crops such as banana, maize and cassava because of fertile soils and high rainfall. Agriculture in the region can be characterized as subsistence farming with complex mixed cropping. Areas above 2500 m.a.s.l. are excluded from the study area as these are deemed unsuitable for banana production. The study area is delimited and stratified using administrative boundaries. For Tanzania, Kenya and Uganda the boundaries correspond to district boundaries, for Rwanda to the prefecture boundaries and for Burundi to the provincial boundaries (derived from Hijmans et al., 2013). For simplicity we will refer to the administrative regions as districts in the remainder of this article.

2.2. Farm survey

The disease incidence dataset was collected by the Crisis Crop Control Program (Abele et al., 2007). This program carried out a large scale survey in banana and cassava growing areas in Burundi, Democratic Republic of Congo (DRC), Kenya, Rwanda, Tanzania and Uganda. A total of 2871 farms were surveyed between July 2006 and July 2007. The survey resulted in a database with information on household characteristics, crop yields and food security. BWB incidence was not measured directly in the field but a proxy related to yield loss was included. Farmers were asked to estimate the yield loss due to BWB as a percentage of their total banana yield. Farm locations were recorded using a handheld GPS. From the overall survey, 1350 farms were located within the study area and included banana cultivation. The spatial distribution of these farms (Fig. 1) clearly shows that sampling took place along the major routes using a convenience sample. Descriptive statistics of disease incidence are presented in Table 1. Almost a third of the selected farms were infected with BWB, with some experiencing a 100% yield loss (Fig. 2). However, large regional differences were found. In Uganda, almost 70% of the farmers experienced yield losses. For the other countries yield losses were much smaller, with average losses not exceeding 10% and with median scores of 0%. Locally, however, the losses were considerable and frequently exceeded 50%.

The survey included several of the commonly occurring limitations to disease mapping described in the Introduction: a convenience sample where a proxy of disease incidence in terms of yield loss was used and that presents a non-normal distribution and high short-distance variability. Despite some of the characteristics of the dataset, the conventional approach of describing the spatial

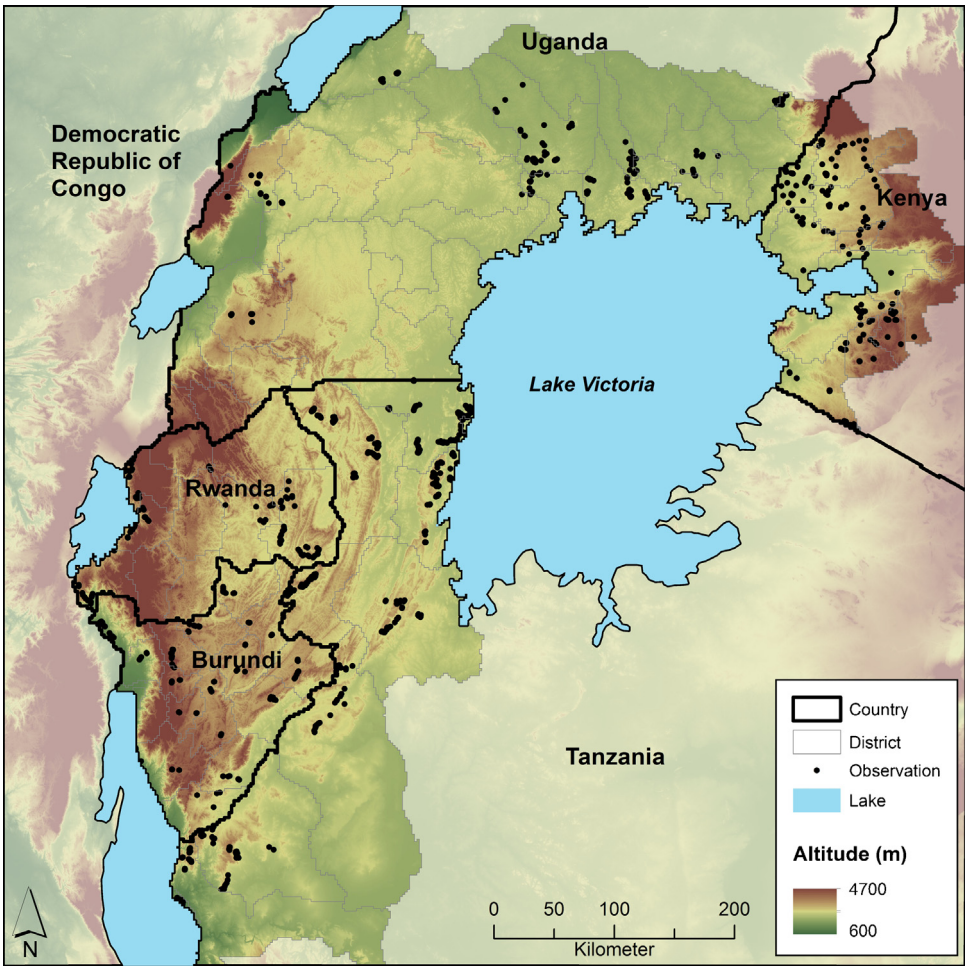

Fig. 1. Map of the East African highlands and altitude. Black dots are observation locations of BWB observations.

Table 1
Descriptive statistics of BWB incidence (expressed in% yield loss) in the East African highlands, where T0 (presence/absence with threshold at 0%) and T25 (severe/not severe with threshold at 25%) are the proportions of observations.

Country	*N*	Mean	Std. dev.	Median	Min.	Max.	Skewness	T0	T25
Burundi	135	6.7	19.6	0.0	0.0	100.0	100	0.26	0.08
Kenya	213	2.8	9.9	0.0	0.0	80.0	80	0.13	0.04
Rwanda	125	2.0	9.5	0.0	0.0	67.0	67	0.08	0.03
Tanzania	548	7.6	20.1	0.0	0.0	100.0	100	0.16	0.13
Uganda	329	20.8	19.9	25.2	0.0	97.0	97	0.72	0.39
Study area	1350	8.0	15.8	5.0	0.0	89.0	89	0.29	0.16

distribution and incidence of crop diseases is to take the arithmetic mean of observations located in an area delineated with administrative boundaries (Tripathi et al., 2009; Bouwmeester et al., 2010; Legg et al., 2006; Night et al., 2011; Tusheremereirwe et al., 2006).

2.3. Mapping disease incidence

We propose to use indicator regression kriging to deal with some of the specific characteristics of the survey data. Indicator regression kriging is a geostatistical interpolation method that can spatially interpolate a binary response variable, making use of point observations of the target variable and auxiliary data (Goovaerts, 1997; Hengl et al., 2004; Webster and Oliver, 2007). The methodology to map disease incidence involves four stages. In the first stage, exploratory data analysis is done and the disease incidence observations are transformed into binary variables. The second stage regresses these binary variables to auxiliary environmental covariates (e.g., terrain, land cover, agro-ecology). These auxiliary covariates must be available as maps, i.e. they must be spatially exhaustive. Since the dependent variable is binary, logistic regression is used. In the third stage the regression model is applied to create an initial map of disease incidence. Finally, in the fourth stage, the regression residuals are interpolated using kriging and added to the regression maps to further improve the spatial prediction of disease incidence.

2.3.1. Data transformation

In general, data transformations aim to transform the data such that they are suitable for a particular statistical technique. In this case, BWB incidence presented a large number of zeros in combination with a very skewed distribution (Fig. 2). Such distributions cannot easily be transformed to normality. Therefore, the variable was transformed into two binary variables. The first variable (T0) represents the presence/absence of BWB, while the second vari-

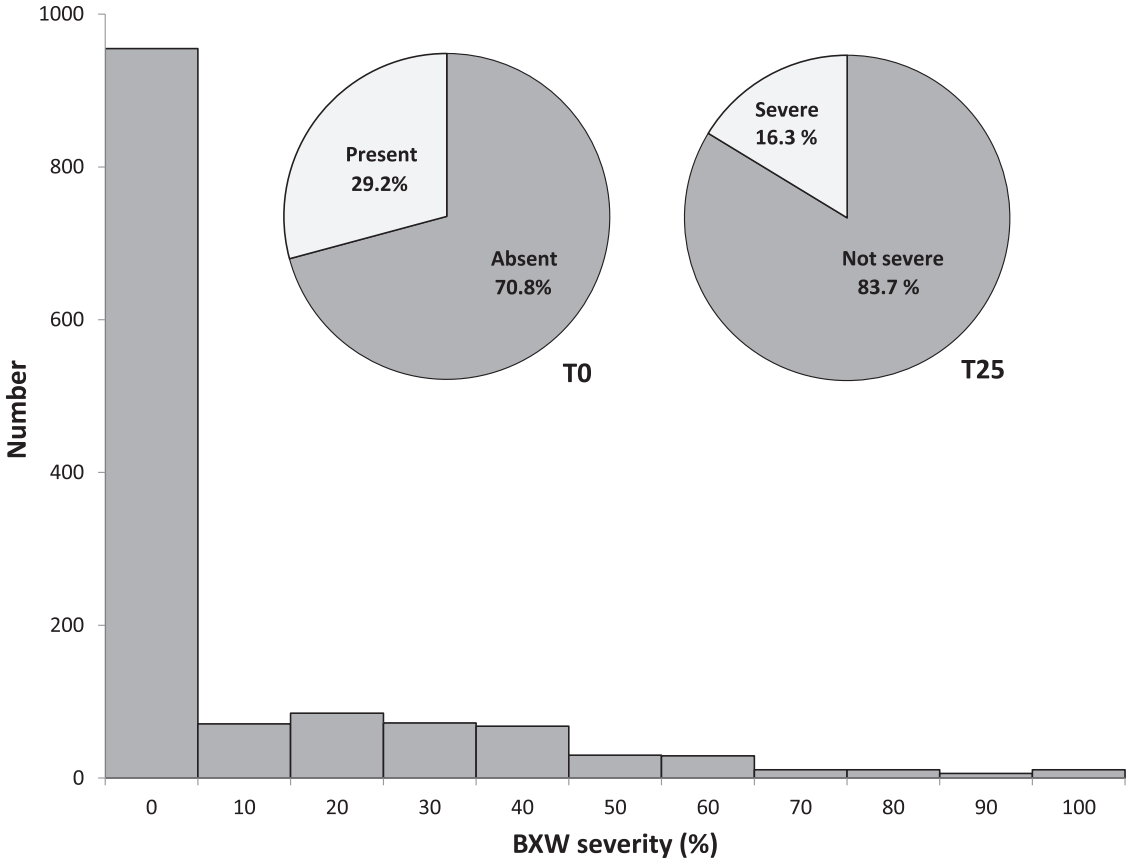

Fig. 2. Histogram of original BWB incidence value and pie charts of binary variables T0, indicating presence/absence of BWB, and T25, indicating a BWB severity of 25% or greater.

able (T25) represents the severity of the disease, being higher or lower than 25%. An, admittedly arbitrary, threshold of 25% was chosen under the assumption that crop yields are reduced to a 'severe' level above this threshold. For disease management and possible policy intervention, T0 and T25 tell two different stories: T0 provides important information on the spatial distribution of the disease whereas T25 indicates the direct impact of the disease on crop yields.

2.3.2. Regression analysis

Regression analyses predicts T0 and T25 at unobserved locations based on their relationship with environmental characteristics. In total, 13 environmental variables were selected (Table 2) that were thought to have a plausible and significant relationship with BWB and for which maps were available in the public domain. These environmental maps function as so-called 'covariates'. The selection of the covariates was based on literature data and expert knowledge. Biruma et al. (2007) reported a relationship between BWB and altitude and precipitation. The spatial pattern of the observations suggest a relationship between BWB and latitude and longitude, while also similar values appear to be spatially clustered. Vegetation cover is derived from satellite images in terms of the Normalized Difference Vegetation Index (NDVI). NDVI and temperature were selected as covariates because these influence the habitat conditions of the host and disease. In addition, the squared temperature, precipitation and vegetation cover were included to account for potential nonlinear relationships.

An overlay of the survey locations and covariates resulted in a database with disease incidence and covariates that serve as input for fitting the logistic regression models. We only briefly review the theory and practical application of logistic regression because these are well-described in the literature (e.g., Hosmer and Lemeshow, 2000).

The outcome of a binary response variable is either 0 or 1. The probability of the outcome being 1 is characterized by p, where $0 \leq p \leq 1$. Logistic regression assumes that the logit transform of p is a linear function of the covariates:

$$\text{logit}(p) \equiv \log\left(\frac{p}{1-p}\right) = \sum_{i=0}^{m} \beta_i \times x_i = x^T \times \beta \quad (1)$$

where m is the number of covariates, T denotes transpose, and where x is a vector of covariates and β a vector of regression coefficients. Note that, x_0 is taken as unity so that β_0 is the intercept of the regression. The regression coefficients are estimated using the maximum likelihood principle. The significance of the logistic regression model can be expressed with the deviance statistic, which is calculated by comparing the log-likelihood of the fitted model with that of the saturated model (i.e., the model that contains as many parameters as there are observations and thus perfectly reproduces the observations):

$$D = -2\log\left(\frac{\text{likelihood fitted model}}{\text{likelihood saturated model}}\right) \quad (2)$$

The ratio in Eq. (2) is the so-called likelihood ratio. The deviance statistic D is a goodness-of-fit measure, with low values of D indicating a better fit. The goodness-of-fit of competing models can now be compared by assessing the change in deviance as a result

Table 2
Environmental covariates used for the regression analysis of BWB incidence in the East African highlands.

Variable	Description	resolution	source
ALT	Altitude above sea level (m)	3 arc s	http://srtm.csi.cgiar.org
PREC	Mean annual precipitation (mm)	30 arc s	www.worldclim.org
PRECSQ	PREC squared (mm^2)	30 arc s	www.worldclim.org
PRECMIN	Mean precipitation of driest month (mm)	30 arc s	www.worldclim.org
PRECVAR	Precipitation seasonality (i.e., coefficient of variation, mm)	30 arc s	www.worldclim.org
LONG	Longitude position (m)		–
LAT	Latitude position (m)		–
NDVI	Mean monthly Normalized Differential Vegetation Index (derived from SPOT images over period 2001–2005) (–)	30 arc s	http://www.spot-vegetation.com/
NDVISQ	NDVI squared (–)	30 arc s	www.worldclim.org
TEMP	Mean annual temperature (°C)	30 arc s	www.worldclim.org
TEMPSQ	TEMP squared ($°C^2$)	30 arc s	www.worldclim.org
TEMPMIN	Mean temperature of coldest month (°C)	30 arc s	www.worldclim.org
TEMPVAR	Temperature seasonality (i.e., coefficient of variation, °C)	30 arc s	www.worldclim.org

Table 3
Estimates and standard errors of coefficients of the logistic regression models for T0 and T25. Interactions are indicated with the ':' sign.

	T0		T25	
	Coefficient	SE	Coefficient	SE
Intercept	2.49E + 02	7.49e + 01	−5.06E + 03	2.31E + 03
LAT	−5.62E − 06	3.16E − 06	5.23E − 04	2.38E-04
ALT	−1.00E − 01	2.77E − 02	–	–
TEMP	−1.67E + 00	4.66E − 01	4.75E + 01	2.16E + 01
TEMPSQ	2.97E − 03	8.79E − 04	−1.12E − 01	5.02E − 02
TEMPVAR	2.41E − 03	7.90E − 04	−3.33E − 03	1.61E − 03
TEMPMIN	1.06E − 01	1.99E − 02	1.23E − 01	2.72E − 02
PREC	1.82E − 02	3.21E − 03	1.34E − 02	3.96E − 03
PRECSQ	−7.38E − 06	1.18E − 06	−5.05E − 06	1.49E − 06
PRECMIN	–	–	−6.65E + 00	2.07E + 00
PRECVAR	–	–	5.38E − 02	2.56E − 02
LAT:ALT	7.58E − 09	2.44E − 09	–	–
LAT:TEMP	–	–	−4.92E − 06	2.23E − 06
LAT:TEMPSQ	–	–	1.16E − 08	5.19E − 09
LAT:PRECMIN	–	–	3.18E − 07	1.08E − 07
ALT:TEMPSQ	5.40E − 07	1.74E − 07	–	–
TEMP:PRECMIN	–	–	3.49E − 02	1.50E − 02
TEMPSQ:PRECMIN	–	–	−8.65E − 05	3.58E − 05

of including or excluding a covariate, using the so-called likelihood ratio test (Hosmer and Lemeshow, 2000):

$$G = D(\text{model without covariate}) - D(\text{model with covariate}) \quad (3)$$

Under the null hypothesis that the regression coefficient of the additional covariate is zero, the likelihood ratio test statistic G is Chi-square distributed. Thus, additional covariates will only be included if this leads to a statistically significant increase in G.

To apply a logistic regression model to T0 and T25 using the binary variables from the 1350 farmers and the covariates given in Table 3, we used the following procedure:

1 Fit a univariate logistic regression model between the response variable (i.e., T0 or T25) and each individual covariate. The significance of each model is assessed with the Wald statistic, which compares the estimated coefficient to its standard error (Hosmer and Lemeshow, 2000):

$$W = \frac{\hat{\beta}}{\mathrm{SE}\left(\hat{\beta}\right)} \quad (4)$$

2 Any covariate whose Wald statistic has a lower P-value than 0.25 is deemed significant and kept when fitting the multivariate model. Using a more conservative level (e.g., 0.05) often fails to identify covariates known to be important (Mickey and Greenland, 1989).

3 Fit a multivariate logistic regression model using all covariates that were significant in the univariate logistic regressions. Next use stepwise regression to reduce the number of covariates, by removing covariates that are not significant at the $P = 0.05$ level, one at a time.

4 Include two-way-interactions between covariates to check for combined effects that improve the likelihood ratio. Initially, all possible covariate interactions are examined. All interactions with a P-value less than 0.05 are deemed significant and included in the model, again using a stepwise approach.

5 The goodness-of-fit of the final model is assessed in terms of deviance and compared to the null model (i.e., the model without covariates) using the likelihood ratio test.

2.3.3. Spatial interpolation of regression residuals

In most cases the regression model will only describe part of the variation in the response variables, for a number of reasons:

1. Not all underlying processes that cause spatial variation in the disease incidence are known (e.g., mode of transmission, susceptibility of the host plant).
2. Some processes may be known to be important but cannot be represented by covariates because these are not available (e.g., cultivar distribution, farm management).
3. Environmental covariate maps representing underlying processes suffer from errors or do not have the appropriate scale (e.g., banana cultivation areas).

In regression kriging, the regression residuals are interpolated with kriging and used to correct the estimate of the regression model. Kriging predicts at unobserved locations by taking a weighted average of the surrounding observations, where the kriging weights depend on the spatial correlation between the variable at the prediction and observation locations. The spatial correlation is characterized by the semivariogram that plots the semivariance, i.e., a measure of the degree of variation, as a function of geographical distance (Goovaerts, 1997). In regression kriging, the residuals from the regression analysis are used instead of the observations directly. We estimated the semivariogram model parameters with restricted maximum likelihood using the GSTAT package in R (Pebesma, 2004). The semivariogram model shape and range were selected from 30 candidate semivariograms, using spherical, exponential and circular models at different range values. The semivariogram model with the largest likelihood was selected. Next, simple kriging of the regression residuals was applied since the regression residuals have zero mean (Webster and Oliver, 2007):

$$\hat{Z}_{\mathrm{SK}}(x_0) = \sum_{i=1}^{n} \lambda_i \times Z(x_i) \quad (5)$$

where $\hat{Z}_{SK}$ is the predicted value of the regression residual at prediction location x_0, the λ_i. are kriging weights, and where the $x_i\,(i = 1, \cdots, n)$ are observation locations. In our study, the residual stands for the observation minus the outcome of the regression model, i.e. Z = T0-$\hat{p}$ or Z = T25-$\hat{p}$. The interpolated regression residuals are now added to the predictions from the logistic regression model. In this study, the spatial resolution of all covariate maps was converted to 2 km × 2 km.

2.4. Spatial aggregation

The high resolution prediction maps made with indicator regression kriging may be too detailed and difficult to translate into clear-cut policy decisions or recommendations. Instead, maps that show the severity of BWB per district could be used to allocate agricultural extension work and to target disease management policies to those districts that need it most. As argued before, such maps are usually made using conventional methods of arithmetic averaging of observations within districts, which ignores the limitations set by the spatial configuration of the observations (e.g., unevenly distributed and clustered observations). These limitations can be avoided by aggregating the predictions obtained with indicator regression kriging for all cells within a district, by calculating:

$$p_d = \sum_{i=1}^{N} p_i \times 100\% \tag{6}$$

where p_d is the aggregated district value, N is the number of cells in the district and p_i the predicted values of cell i of the district. Thus, p_d is the average value for all cells in a district and represents the percentage of the area in the district with disease incidence above the threshold. This aggregated average may be compared to the 'naive' averages calculated with the conventional method. It is expected that the resulting values will be different as the conventional method takes account of the observations only whereas the geostatistical method uses covariates and spatial autocorrelation.

2.5. Accuracy of predictions

The accuracy of the predictions is assessed through a cross-validation. For the conventional method the prediction error is determined by calculating the arithmetic mean of each district as described before, with the difference that now each observation is one by one removed from the dataset before the mean is calculated. The value predicted for the observation location can then be compared to the value that was actually observed. In seven districts only one observation was located. These observations were excluded from the validation because no arithmetic mean could be calculated for those districts if the observation was removed. For the geostatistical method the prediction error is determined in a similar way. A prediction is made for each observation location without including the observation itself in the kriging procedure. The value predicted can then be compared to the observed value. Accuracy metrics that are computed from the comparison are the Mean Error and Root Mean Squared Error.

3. Results

3.1. Exploratory data analysis and conventional mapping

The summary statistics in Table 1 and Fig. 2 reveal that BWB incidence has a very skewed distribution with many cases where no BWB was found. Fig. 3a and b show the results of applying the conventional approach of calculating district averages of T0 and T25 for the study area. From these figures it appears that:

- many districts are disease free, as none of the interviewed farmers reported presence of BWB,
- for many districts no predictions are made, as no farmers were interviewed in these districts,
- the disease appears to be spatially clustered, where adjacent districts often have similar disease values (although occasionally adjacent districts have large differences in disease values), and
- although T0 and T25 show similar patterns, there is quite a difference in Burundi, where the disease is widespread but where severe yield losses occur in only a relatively small number of districts.

3.2. Regression analysis

A univariate logistic regression model was fit for T0 and T25. All covariates were highly significant for T0 (all $P < 0.00$) and included when fitting the model. For T25 also all covariates were included, although overall the significance was lower (all $P < 0.24$). The covariates themselves were often strongly cross-correlated. Not surprisingly, the correlation was highest for the covariates that were based on the same variable, such as the mean and minimum annual temperature ($r = 0.91$, $P = 0.00$). The correlation was also high between precipitation variability and vegetation index ($r = -0.76$, $P = 0.00$), but also for less obvious covariates, such as precipitation variability and latitude ($r = -0.86$, $P = 0.00$). The initial multivariate logistic regression models for T0 and for T25 included all 13 covariates, but only the significant ones were kept. As explained in Section 2.3.2, two-way-interactions that were significant were added to the models to further improve the likelihood ratio. The final models are presented in Table 3. The deviance of the T0 model is 27.9% smaller than for the null model whereas the deviance of the T25 model is 23.0% smaller. This indicates that the covariates in both regression models explain a larger portion of the variation than the null model. It also indicates that the T0 model explains more of the spatial variation than the T25 model. Both models are highly statistically significant, partly as a result of the large sample size. The regression maps for T0 and for T25 show a similar spatial pattern with high values in the North-West and low values predominately in the South. The similarity can partly be explained by the fact that the input data of T0 and T25 are largely identical, since only the observations differ where BWB is present but its severity is lower than 25%. Noticeable is the 'pepper and salt' structure, which results from the highly diverse environment in the region and the resulting large spatial variation in covariate maps over short distances.

3.3. Simple kriging of regression residuals

Semivariograms were calculated for the regression residuals of T0 and T25 using restricted maximum likelihood (Fig. 4). For both cases, the highest likelihoods were obtained using circular semivariogram models. For T0 a moderate spatial dependency was present with a nugget to sill ratio of 50%. The spatial correlation for T25 was lower with a nugget to sill ratio of 62%. The nugget indicates high variation between residuals at close distance. The range represents the maximum distance over which the residuals are still correlated and was 14 km for T0 and 12 km for T25. The spatial dependence of the regression residuals indicates presence of spatial patterns caused by environmental factors that were not included in the regression analysis.

3.4. Prediction maps

The regression map combined with the interpolated residuals results in the final prediction maps for both thresholds (Fig. 3c and d). The predictions can be interpreted as the probability that a cell

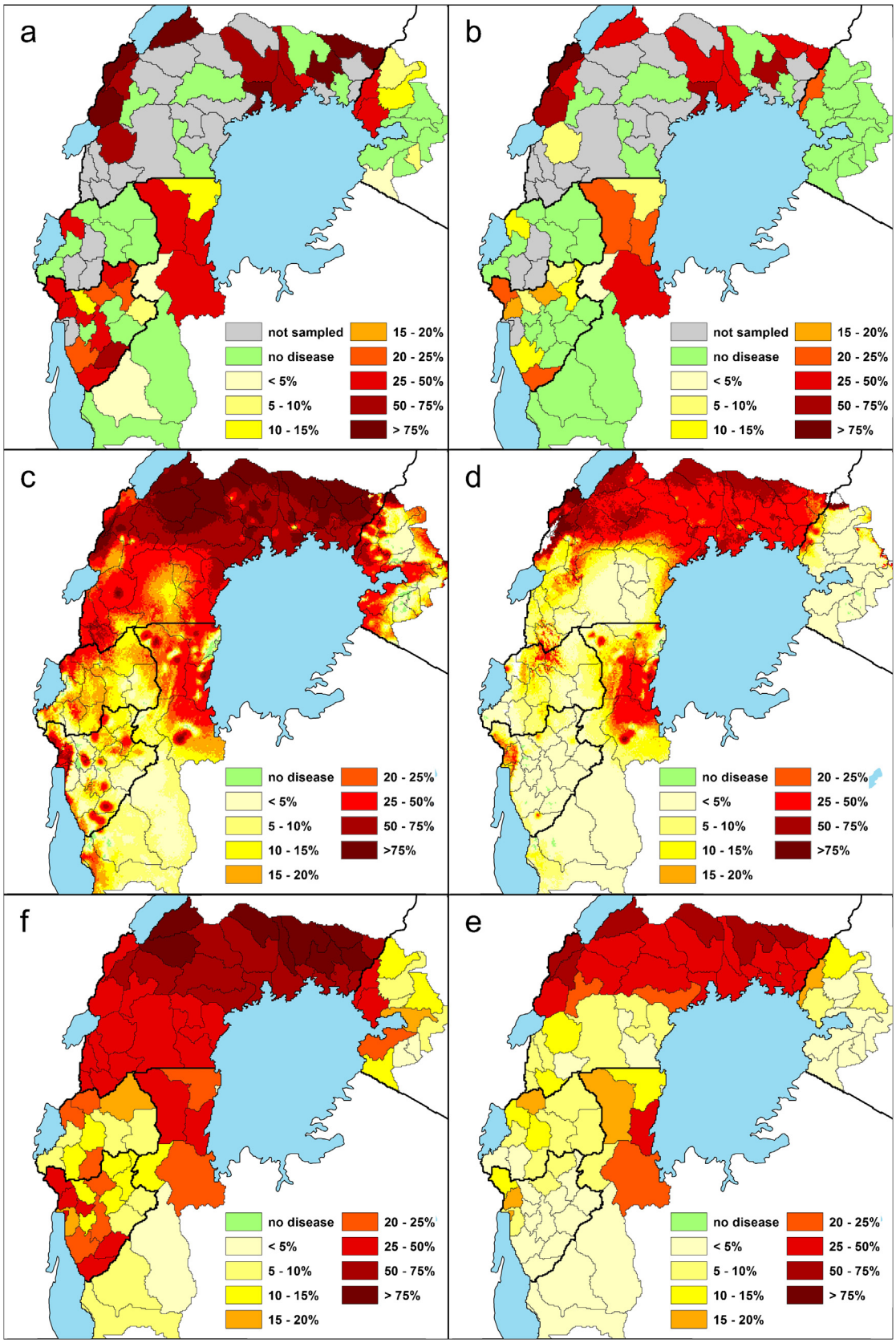

Fig. 3. Maps of BWB in the East African highlands as arithmetic mean per district in terms of T0 (a) and T25 (b), as regression kriging predictions in terms of T0 (c) and T25 (d), and as proportion of district in which BWB is present (e) and proportion of district in which BWB exceeds the 25% threshold (f).

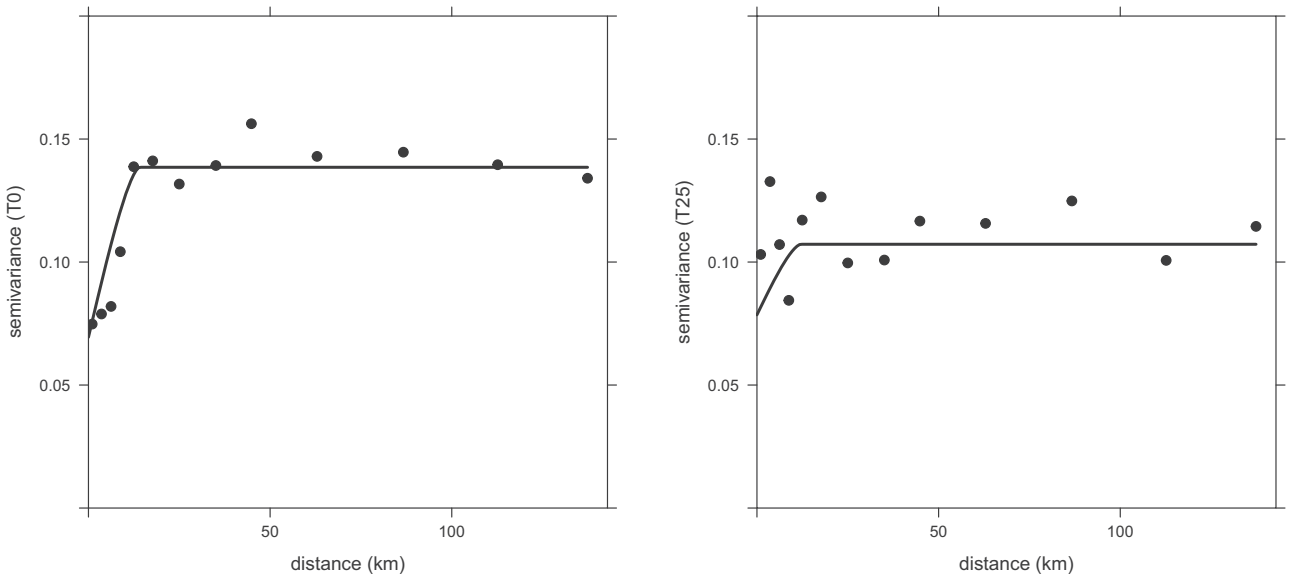

Fig. 4. Semivariograms for BWB for T0 (a: nugget = 0.069, sill = 0.138, range = 14 km) and T25 (b: nugget = 0.079, sill = 0.127, range = 12 km) for the East African highlands.

exceeds the threshold. For example, a T0 prediction value of 40% indicates a 40% probability that farms in the cell are infected with BWB. For T25, the predictions can be similarly interpreted as the probability that a cell exceeds the 25% threshold. The patterns on both maps are similar, though as expected the values on the T25 map are generally lower. The influence of regression is obvious as the general pattern resembles that of the regression map alone. Noticeable are the circular shaped patches of higher and lower values. These result from the kriging, are centered around observations and are most pronounced when observation values are markedly different from the regression predictions. The patches are circular and not very large because the relatively small range of the semivariogram limits the zone of influence of observations. In principle, the further away from an observation the less influence it has and the more the covariates take over.

3.5. Spatial aggregation

The prediction maps of combined regression and interpolation (Fig. 3c and d) are calculated for each individual cell. These high resolution maps were aggregated to 83 districts (Fig. 3e and f). These maps can be interpreted as the proportion of the districts in which the thresholds are exceeded. The map of T0 depicts the proportion of each district that is infected with BWB. The map of T25 depicts the proportion of each district with a severe yield loss (>25%). Both maps show a similar spatial pattern. On the T0 map, BWB seems to be concentrated in Northern Uganda, where in most districts more than 50% of the area was infected, often at a severe level. In contrast, the South-West and East seem to be relatively better off, as BWB is present everywhere but predominantly at low levels.

3.6. Accuracy of predictions

The accuracy of the predictions are quantified using cross-validation. The calculated differences between observed and predicted values for both methods are shown in Fig. 5. As expected, the mean error for both methods is close to zero. The Mean Error for the geostatistical method is 9.6×10^{-4} for T0 and 8.7×10^{-4} for T25. For the conventional method the Mean Errors are smaller, namely 4.4×10^{-11} for T0 and 3.0×10^{-15} for T25. In any case, the Mean Errors are negligibly small compared to the Root Mean Squared Errors, which are 0.30 for T0 and 0.30 for T25 for the geostatistical method, against 0.36 for T0 and 0.34 for T25 for the conventional method. Note also that the cross-validation errors for the geostatistical method are clearly centered around a single mean value, whereas those for the conventional method appear to have several peaks.

4. Discussion

4.1. Proxies instead of measurements of disease incidence

A serious limitation of many surveys is that they do not directly measure disease incidence but instead rely on proxies which could be related to e.g., the occurrence of the pathogen or yield loss. This was also the case for this study on BWB incidence. The survey was recorded on the basis of a farmer recall of yield loss due to BWB. The advantage of the use of proxies is that with the same amount of resources the number of observations can be larger compared to surveys where the disease incidence is actually measured. The results of the survey were interpreted in this study through two indicators T0 and T25. The methodology that is being proposed allows for a useful evaluation of these two indicators but does not provide information on the quality of the proxies and subsequent indicators. To properly asses the quality of the patterns of BWB in the East African highlands the proxies need to be validated.

4.2. High resolution prediction maps

The methodology allows to map patterns of BWB in the East African highlands at a relatively high resolution using indicator regression kriging, despite limitations of the initial survey dataset. Both T0 and T25 show similar disease patterns, indicating that BWB is a serious problem in the north of Uganda. In all other countries BWB is also present but mostly at lower levels. Both disease pattern maps are to a large extent the result of the regression analysis, as becomes apparent when comparing the regression maps with the prediction maps. It is hard to quantify precisely how much of the prediction is determined by regression and how much by kriging as the ratio varies throughout the study area. Overall, regression explains more of the variance of the observations than kriging, as the standard deviation of the regression values for T0 is 0.27 and the standard deviation of kriging values 0.06. Convenience sampling resulted in large areas that were not sampled. The semivariogram

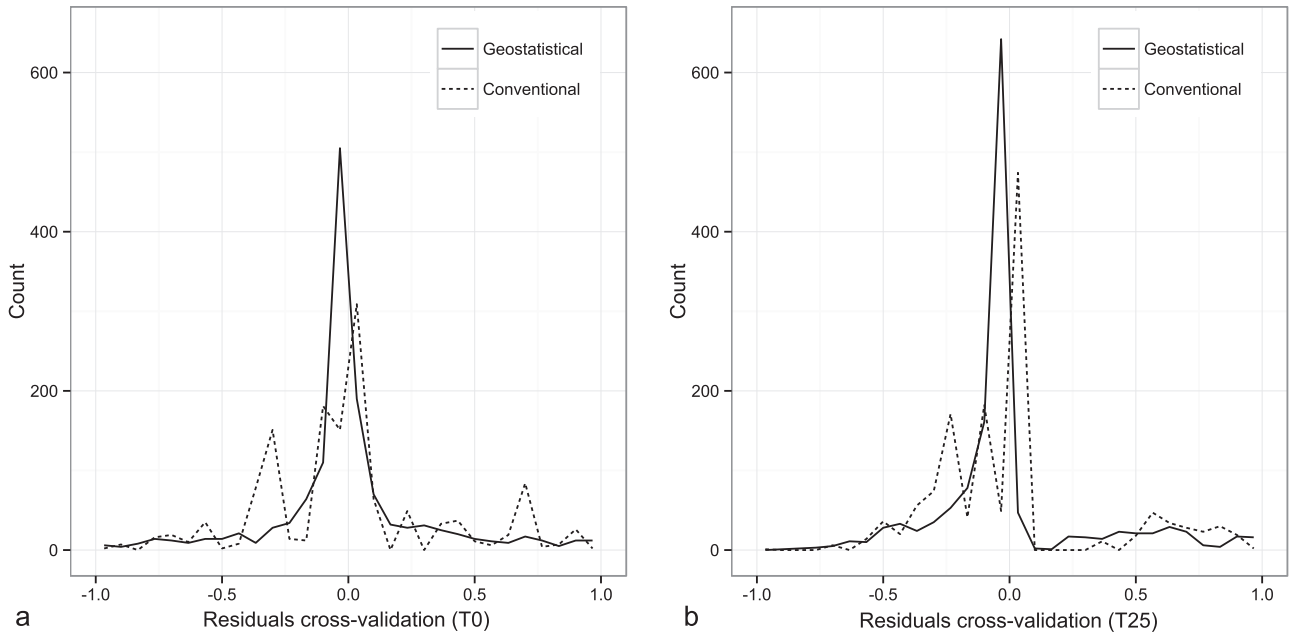

Fig. 5. Differences calculated by cross-validation of predicted and observed values of BWB in the East African highlands. Differences calculated with geostatistical and conventional methods for T0 (a) and T25 (b).

of the residuals of the logistic regression had a relatively short range of 14 km for T0 and 12 km for T25. This makes it practically impossible to interpolate towards areas with no observation points. In those areas, the predictions are entirely determined by the regression surface. The kriging effect is shown by the circular patches on the maps, with lower and higher values than the regression maps.

4.3. Spatial aggregation of high resolution maps

The prediction maps (Fig. 3c and d) show that BWB is widespread at varying intensity. Spatial aggregation of these maps (Fig. 3e and f) allows policy makers to prioritize, target disease management policies and allocate agricultural extension work. For example, a policy maker could attempt to manage BWB by intervening once a certain proportion of a district has exceeded a severe level. Alternatively, a policy maker could contain the spread of the disease by targeting districts that have low levels of BWB but border districts with high levels. Obviously many other control strategies are possible as well as different threshold maps, where in this study only T0 and T25 were chosen.

4.4. Comparison of aggregated and conventional maps

The spatially aggregated maps (Fig. 3e and f) can be compared to the maps made with the conventional method of averaging samples (Fig. 3a and b). A number of differences become apparent as a result of the comparison:

- The aggregated maps cover all 83 districts in the study area, while the conventional maps cover 62 districts only. This is because indicator regression kriging predicts everywhere, while conventionally only values are predicted for sampled districts. Policy makers could benefit from the additional information for 21 districts.
- The conventional maps (Fig. 3a and b) indicate that BWB is not present in 24 out of 62 districts. In contrast, the aggregated maps (Fig. 3e and f) predict BWB to be present in all districts. There are even some districts in Uganda where the conventional map indicates that BWB is absent, while the aggregated maps predict the disease to be widely spread and in some cases even exceeding the severe level. This is because all responses in those 24 districts were negative, while positive responses in neighboring districts by definition do not influence the conventional maps. However, the aggregated maps get positive scores through regression and because kriging interpolated across district borders. This result could make policy makers realize spatial spread of BWB is not restricted by administrative borders.
- The aggregated maps present a more gradual pattern than the conventional maps. This pattern results to a large extent from the regression of mostly smooth covariates. It also results from the kriging, as observations are clustered and the influence of a cluster decreases when moving away from it.

The large differences in maps made with the same data but with different methods (i.e., conventional and geostatistical) underline the importance of the selection of the methodology. The differences between the methodologies for the sampled 62 districts are presented in Fig. 6. From the figure it appears that the larger the number of observations within a district the more the predicted district values resemble those calculated by conventional averaging. Indeed the Spearman rank correlation between the number of observations per district and the absolute differences in predictions using conventional and geostatistical approaches was negative ($r = -0.52$ in the case of T0 and $r = -0.17$ in the case of T25), although only in the case of T0 was it significantly different from zero (P-value $= 2.0 \times 10^{-5}$).

4.5. Quality of high resolution and aggregated maps

Accurate maps on disease patterns are of high importance. Indicator regression kriging is able to deal with some of the specific limitations of the survey data. The cross-validation (Fig. 5) gave some idea about the quality of the predictions. It showed that the Root Mean Squared Error of the geostatistical methodology is slightly smaller than that obtained with the conventional method. The cross-validation results should be interpreted with care because the observations were clustered, implying that the so-obtained accuracy measures may not be a realistic assess-

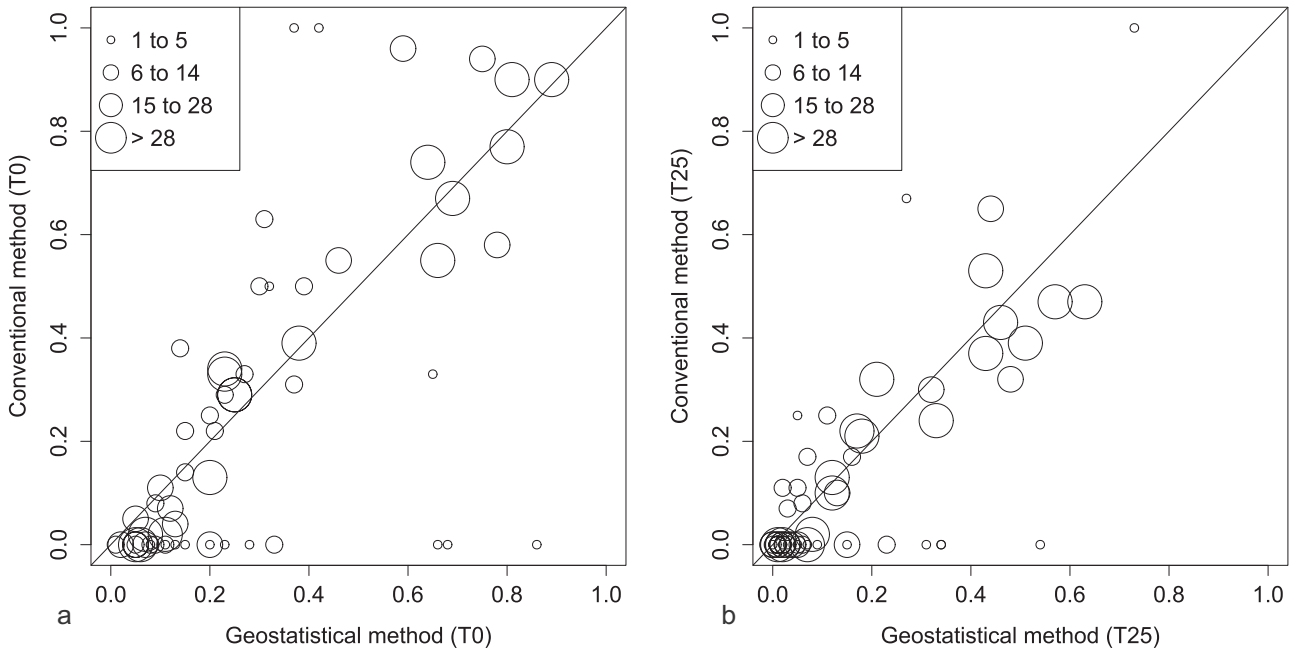

Fig. 6. Scatter plots of aggregated predictions of BWB disease per district in the East African highlands in terms of T0 (a) and T25 (b). Predictions obtained with the geostatistical method on the *x*-axis, with the conventional method on the *y*-axis. Size of open circles increases with the number of surveyed farms located within the district.

ment of the true map accuracy. Unfortunately, no independently collected validation data were available to validate the maps. Therefore we recommend collecting independent data for validation in future, preferably with locations selected using probability sampling. Cross-validation provides useful information, but it cannot compete with a truly independent, design-based validation (e.g., Knotters and Brus, 2013).

4.6. Limitations of the dataset

The dataset was not collected according to a well-defined sampling design, but instead clustered with large spatial gaps (e.g., preferential or convenience sampling). Where the clusters were situated the kriging had a marked influence on the final predictions, while further away from observation locations regression took over. In principle, indicator regression kriging can handle these type of datasets, but more accurate maps would probably be generated if the observations would be more uniformly distributed (Stein and Ettema, 2003). However, a uniformly distributed sampling design would have practical complications as the study area is very large and some parts are difficult to access. As we have shown, maps of covariates partly fill in the gaps that result from the clustered sampling pattern, and thereby decrease the need for regular sampling. Additional research is recommended to exploit relationships between disease incidence and environmental variables, in an effort to identify additional and more informative environmental maps that could then act as covariates.

A large short distance variability in observations was found, which could be incorporated in future survey designs. Observations at larger support (i.e., the area or volume over which a measurement is made) would possibly resolve this issue, as local variation is removed when observations are averaged over larger areas (Goovaerts, 1997).

No standard diagnostics were used to assess BWB, instead it was determined on farmer recall. This may introduce measurement error as an expected estimate of yield loss is difficult to make. This was also shown by a study on a cassava disease in Burundi and Rwanda (Bouwmeester et al., 2012), that used three datasets for mapping and concluded that results between resulting maps were quite different, despite the fact that the datasets considered the same disease. Measurement error could perhaps be reduced by using standard diagnostics and clear and unambiguous instructions.

4.7. Limitations of the methodology

The variable of interest could not be transformed to normality because of the many zeros and odd distribution, while ordinary and regression kriging are tailored to the normal distribution (Hengl et al., 2004; Diggle and Robeiro, 2007). Our solution was to use indicator kriging, a widely accepted and applied method that has proven useful in many applications, also in our case. However, there has been criticism on the method, especially on regression indicator kriging (Papritz, 2009). A statistical more elegant and robust method is Generalized Linear Geostatistical Modelling (Diggle and Robeiro, 2007). However, practical application of this methodology is cumbersome and computationally intensive, and it is unclear whether it can adequately deal with dependent variables that exhibit strong spatial dependence (Kempen et al., 2012). From a practical application point of view indicator regression kriging is preferred, but it is worthwhile to consider alternative approaches and compare methods and results using independent validation.

4.8. Implications for disease mapping and control

Efficient use of resources to target investments in the extension of disease management and contingency measures to avoid the spread of diseases rely on proper disease mapping. There has been significant research on optimal sampling designs of surveys (e.g., Stein and Ettema, 2003; De Gruijter et al., 2007), but practice learns that for various reasons nearly all surveys have limitations. This is particularly true for disease mapping, where infestation levels or impacts are difficult to observe. In addition, we often see significant short-distance variability. The current paper shows how geostatistical techniques can be used to overcome some of these problems. This allows for a better use of existing surveys or of surveys that are collected for different purposes. However, we also showed that relatively simple adaptations to surveys may improve results, e.g.,

by making sure that observations are done under a broader range of environmental circumstances (to allow for a better derivation of correlations with auxiliary information) or a reasonable coverage of the study area (to provide a better basis for interpolation). A general problem for the interpretation of survey results remains if crop diseases are observed through proxy variables like yield loss, the occurrence of symptoms, or, like in this study, the estimated yield loss. A verification of the relation between the proxy variable and the actual disease incidence is required for a proper interpretation of the resulting disease maps and their quality.

In all cases it is important to study the survey data and identify the limitations before subsequent analysis, like spatial interpolation or aggregation. By recognizing the limitations, the subsequent analysis can be adapted to deal with these and to develop best possible disease maps. These maps can subsequently be used for contingency, extension of disease management and for yield forecasting (El Jarroudi et al., 2012).

5. Conclusions

The design of regional surveys on disease spread is often suboptimal for mapping purposes. We apply a method that overcomes some of these problems and generates high-resolution maps of the spatial distribution of crop diseases from regional survey datasets and spatially exhaustive environmental covariates. The maps show that the disease bacterial wilt of banana is widespread on farms throughout the entire study area with locally severe impacts, especially in Uganda. The maps are more accurate than those generated with conventional methods, because environmental maps and spatial correlation are incorporated as covariates, and because the disease incidence is predicted for the entire study area, including districts without observations. The method can deal with imperfect statistical properties and imperfect spatial sampling designs but depends on the quality of proxies that are being used. The high-resolution maps can provide a valuable basis for contingency measures or the extension of proper disease control.

Acknowledgements

The data were collected within the framework of the USAID-funded Crop Crisis Control Program. The contribution of J.J. Stoorvogel was part of the Wageningen University research program “Panama disease: Multi-level solutions for a global problem”.

References

Abele, S., Twine, E., Legg, C., 2007. Food Security in Eastern Africa and the Great Lakes Region. Int. Inst. for Tropical Agric., Ibadan, Nigeria.

Abele, S., Twine, E., Bouwmeester, H., 2008. Food Security in Eastern Africa and the Great Lakes Region in a Nutshell. C3P Food Security Briefs No.8. Int. Inst. for tropical Agric., Ibadan, Nigeria.

AfDB, OECD, UNDP, 2015. African economic outlook. In: Regional Development and Spatial Inclusion. African Development Bank.

Biruma, M., Pillay, M., Tripathi, L., Blomme, G., Abele, S., Mwangi, M., Bandyopadhyay, R., Muchunguzi, P., Kassim, S., Nyine, M., Turyagyenda, L., Eden-Green, S., 2007. Banana Xanthomonas wilt: a review of the disease: management strategies and future research directions. Afr. J. Biotechnol. 6, 953–962.

Blomme, G., Ploetz, R., Jones, D., De Langhe, E., Price, N., Gold, C., Geering, A., Viljoen, A., Karamura, D., Pillay, M., Tinzaara, W., Teycheney, P.Y., Lepoint, P., Karamura, E., Buddenhagen, I., 2013. A historical overview of the appearance and spread of Musa pests and pathogens on the African continent: highlighting the importance of clean Musa planting materials and quarantine measures. Ann. Appl. Biol. 162, 4–26.

Blomme, G., Jacobsen, K., Ocimati, W., Beed, F., Ntamwira, J., Sivirihauma, C., Ssekiwoko, F., Nakato, V., Kubiriba, J., Tripathi, L., Tinzaara, W., Mbolela, F., Lutete, L., Karamura, E., 2014. Fine-tuning banana Xanthomonas wilt control options over the past decade in East and Central Africa. Eur. J. Plant Pathol. 139 (2), 265–281.

Bouwmeester, H., Abele, S., Manyong, V.M., Legg, C., Mwangi, M., Nakato, V., Coyne, D., Sonder, K., 2010. The potential benefits of GIS techniques in disease and pest control: an example based on a regional project in Central Africa. Acta Hortic. 879, 333–340.

Bouwmeester, H., Heuvelink, G.B.M., Legg, J.P., Stoorvogel, J.J., 2012. Comparison of disease patterns assessed by three independent surveys of cassava mosaic virus disease in Rwanda and Burundi. Plant Pathol. 61, 399–412.

Brentu, F.C., Oduro, K.A., Offei, S.K., Odamtten, G.T., Vicent, A., Peres, N.A., Timmer, L.W., 2012. Crop loss, aetiology, and epidemiology of citrus black spot in Ghana. Eur. J. Plant Pathol. 133, 657–670.

CGIAR-Consortium for spatial infrastructure, 2008. shuttle radar topographic mission (SRTM) digital elevation data. [http://srtm.csi.cgiar.org]. (accessed 15.04.09).

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2007. Sampling for Natural Resource Monitoring. Springer, Berlin, Germany.

Diggle, P.J., Robeiro, P.J., 2007. Model-Based Geostatistics. Springer Science–Business Media, New York, USA.

El Jarroudi, M., Kouadio, L., Bertrand, M., Curnel, Y., Giraud, F., Delfosse, P., Hoffmann, L., Oger, R., Tychon, B., 2012. Integrating the impact of wheat fungal diseases in the Belgian crop yield forecasting system (B-CYFS). Eur. J. Agron. 40, 8–17.

FAO, 2012. FAOSTAT Crop Production Data. Food and Agricultural Organization, Rome.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York, USA.

Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120, 75–93.

Hijmans, R., Cameron, S., Parra, J., Jones, P., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.

Hijmans, R., Garcia, N., Wieczorek, J., 2013. Global administrative units V.2.0. [http://www.gadm.org]. (accessed July 2013).

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression. John Wiley & Sons Ltd., Chichester, UK.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., 2012. Soil type mapping using the generalised linear geostatistical model: a case study in a Dutch cultivated peatland. Geoderma 189–190, 540–553.

Knotters, M., Brus, D.J., 2013. Purposive versus random sampling for map validation: a case study on ecotope maps of floodplains in the Netherlands. Ecohydrology 6, 425–434.

Lamichhane, J.R., Fabi, A., Ridolfi, R., Varvaro, L., 2013. Epidemiological study of hazelnut bacterial blight in central Italy by using laboratory analysis and geostatistics. PLoS One 8 (2), e56298, http://dx.doi.org/10.1371/journal.pone.0056298.

Legg, J.P., Owor, B., Sseruwagi, P., Ndunguru, J., 2006. Cassava mosaic virus disease in East and Central Africa: epidemiology and management of a regional pandemic. Plant Virus Epidemiol. 67, 355–418.

Mickey, R.M., Greenland, S., 1989. The impact of confounder selection criteria on effect estimation. Am. J. Epidemiol. 129, 125–137.

Munar-Vivas, O., Morales-Osorio, J.G., Castañeda-Sánchez, D.A., 2010. Use of field-integrated information in GIS-based maps to evaluate Moko disease (*Ralstonia solanacearum*) in banana growing farms in Colombia. Crop Prot. 29, 936–941.

Night, G., Asiimwe, P., Gashaka, G., Nkezabahizi, D., Legg, J.P., Okao-Okuja, G., Obonyo, R., Nyirahorana, C., Mukakanyana, C., Mukase, F., Munyabarenzi, I., Mutumwinka, M., 2011. Occurrence and distribution of cassava pests and diseases in Rwanda. Agric. Ecosyst. Environ. 140, 492–497.

Ozor, N., Umunnakwe, P.C., Acheampong, E., 2014. Challenges of food security in Africa and the way forward. Development 56, 404–411.

Papritz, A., 2009. Limitations of indicator kriging for predicting data with trend. In: D. Cornford, G., Dubois, D., Hristopulos, E. Pebesma and J. Pilz (Eds.), Proceedings StatGIS 2009 (www.stat.aau.at/Tagungen/statgis/2009/).

Park, Y.L., Perring, T.M., Krell, R.K., Hashim-Buckey, J.M., Hill, B.L., 2012. Spatial distribution of pierce's disease related to incidence, vineyard characteristics, and surrounding land uses. Am. J. Enol. Vitic. 62 (2), 229–238.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.

Rakib, M.R.M., Bong, C.F.J., Khairulmazmi, A., Idris, A.S., 2014. Occurrence and spatial distribution of *Ganoderma* species causing upper and basal stem rot in oil palm. J. Food Agric. Environ. 12, 360–364.

Rees, E.E., Ibarra, R., Medina, M., Sanchez, J., Jakob, E., Vanderstichel, R., St-Hilaire, S., 2014. Transmission of *Piscirickettsia salmonis* among salt water salmonid farms in Chile. Aquaculture 428–429, 189–194.

Savary, S., Ficke, A., Aubertot, J.N., Hollier, C., 2012. Crop losses due to diseases and their implications for global food production losses and food security. Food Secur. 4, 519–537.

Spiertz, H., 2012. Avenues to meet food security. The role of agronomy on solving complexity in food production and resource use. Eur. J. Agron. 43, 1–8.

Stein, A., Ettema, C., 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. Agric. Ecosyst. Environ. 94, 31–47.

Thompson, A.H., Narayanin, C.D., Smith, M.F., Slabbert, M.M., 2011. A disease survey of Fusarium wilt and Alternaria blight on sweet potato in South Africa. Crop Prot. 30, 1409–1413.

Tripathi, L., Mwangi, M., Abele, S., Aritua, V., Tushemereirwe, W.K., Bandyopadhyay, R., 2009. Xanthomonas Wilt a threat to banana production in East and Central Africa. Plant Dis. 93, 440–451.

Tusheremereirwe, W.K., Okaasai, O., Kubiriba, J., Nankiga, C., Muhangi, J., Odoi, N., Opio, F., 2006. Status of banana bacterial wilt in Uganda. Afr. Crop Sci. J. 14, 73–82.
Van Ittersum, M.K., Cassman, K.G., Grassini, P., Wolf, J., Tittonell, P., Hochman, Z., 2012. Yield gap analysis with local to global relevance—a review. Field Crops Res. 143, 4–17.
Wang, N., Jassogne, L., van Asten, P.J.A., Mukasa, D., Wanyama, I., Kagezi, G., Giller, K.E., 2015. Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. Eur. J. Agron. 63, 1–11.
Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, 2nd ed. John Wiley & Sons Ltd, Chichester, UK.
Zinga, I., Chiroleu, F., Legg, J.P., Lefeuvre, P., Komba, E.K., Semballa, S., Yandia, S.P., Mandakombo, N.B., Reynaud, B., Lett, J.-M., 2013. Epidemiological assessment of cassava mosaic disease in Central African Republic reveals the importance of mixed viral infection and poor health of plant cuttings. Crop Prot. 44, 6–12.