Query Details

Back to Main Page

There is no Author Query !

Research Article

V Authors' important comment to the publisher:

Please note that we will only give permission to publish after we receive and approve a corrected proof.

alidating models of one-way land change: an example case of forest insect disturbance

Saeed Harati

Email : saeed.harati.asl@umontreal.ca

Liliana Perez

Roberto Molowny-Horas

Robert Gilmore Pontius Jr

Laboratoire de Géosimulation Environnementale (LEDGE), Département de Géographie, Université de Montréal, Montréal, QC, H2V 0B3, Canada

CREAF, 08193, Cerdanyola del Vallès, Spain

School of Geography, Clark University, Worcester, MA, 01610, USA

Received: 14 August 2020 / Accepted: 21 May 2021

Abstract

Context

Validation of models of Land Use and Cover Change often involves comparing maps of simulated and reference change. The interpretation of differences between simulated and reference change depends on the characteristics of the process being studied. Our paper focuses on validation of models of one-way land change processes that spread in space.

Objectives

Our objective is to develop a method for validation of one-way land change models, such that the method provides objective information about the spatial distribution of errors.

Methods

Using distance analysis on reference data, we build a baseline model for comparison with simulations. We then simultaneously compare the four maps of reference at initial time, reference at final time, simulation at final time, and baseline at final time. We also use Total Operating Characteristic curves and multiple-resolution map comparison. We illustrate the methods with a simulation of forest insect infestations.

Results

The methods give insights concerning the reference data, as well as to information concerning the spatial distribution of misses, hits, and false alarms with respect to initial points of infestations. The new methods reveal that the simulations underestimated change near initial points of spread.

Conclusions

The spatial distribution of errors is a topic of land change models that deserves attention. For models of one-way, geographically-spreading processes, we recommend that validation should distinguish between near and far allocation errors with respect to initial points of spread.

Model validation
One-way change

Total operating characteristic
Multiple resolution
Distance analysis
Area partition

Introduction

Validation of land change models

Modelling is a major theme of the Land Use and Land Cover Change (LUCC) science (Lambin et al. 2006; de Sousa-Neto et al. 2018). Models are useful tools in science and policy applications, but before they can be used, their credibility should be evaluated (Pontius Jr et al. 2004, 2008; Pérez et al. 2013). Validation is an important step in the process of evaluation of any model. According to Rykiel (1996), "validation is a demonstration that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model". Measuring a model's accuracy and deciding a satisfactory level have been subjects of debate and disagreement in the scientific community (Verburg et al. 2006). Examples of intended uses of a model are to project future pathways of change and to run tests to understand processes of change and quantify our knowledge about them (Lambin et al. 2006; Brown et al. 2012). It is essential to consider model objectives when validating the model (Batty and Torrens 2005; Brown et al. 2012, 2013; National Research Council 2014; van Vliet et al. 2016). LUCC literature recommends development of four classes of model testing methods: sensitivity analysis to test variation of model results with changing model parameters, uncertainty analysis to test variation of stochastic model results without changing model parameters, structural validation for models built with the aim of understanding change processes, and pattern validation for models aiming at prediction of future changes (Brown et al. 2013; National Research Council 2014). The latter category usually employs techniques of map comparison (Pontius Jr 2000; Foody 2004; Hagen-Zanker 2006; White 2006; Brown et al. 2013; National Research Council 2014). This category of validation methods is further classified into two groups: tests of erroneous composition of land classes i.e. quantity disagreement, and tests of erroneous configuration of land classes i.e. allocation disagreement (Pontius Jr and Millones 2011; van Vliet et al. 2016). A review of 114 models in land change literature found that the most common validation approach among them was location accuracy (van Vliet et al. 2016).

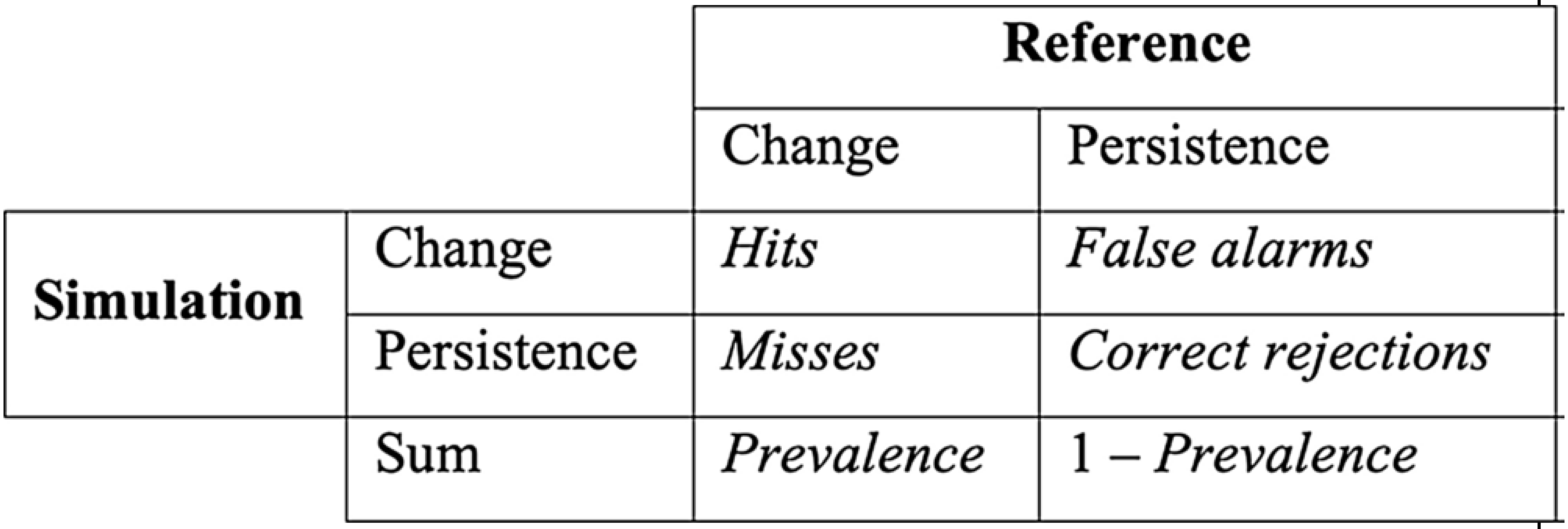
Definitions

Multiple disciplines apply techniques of map comparison and model validation—some examples are urban growth (Pijanowski et al. 2005; Chen and Pontius Jr 2010), landscape ecology (Paudel and Yuan 2012; Cushman et al. 2017), forestry (Rollins et al. 2004), agriculture (Li et al. 2012), conservation (Hermoso et al. 2018), and remote sensing (Foody 2004)—with each discipline using its own terminology. In order to avoid confusion about meanings of the terms used throughout the text, we define them in this section.

The basis of the analyses described in this paper is the contingency table, also known as the confusion matrix or the error matrix. This table summarises the comparison of two binary maps of change. Generally, in model validation, one of the maps shows the change simulated by the model, and the other shows the reference change for comparison. The professional convention is that the rows of the contingency table indicate the simulation change, and the columns indicate the reference change. The value of each cell of the table is the count of units, e.g., pixels, with that cell's row category in simulation and its column category in reference. The sum of values in a row of the contingency table equals the quantity of the respective category in the simulation. Likewise, the sum of values in a column of the contingency table equals the quantity of the respective category in the reference. Finally, the sum of all cells of the contingency table equals the total count of units, e.g. pixels, in the study area. It is also possible to divide all values in the table by this total count, which is what Fig. 1 shows. Then, the values will be proportions of the study area, and their total will be 1.

Fig. 1

Outline of a contingency table



		Reference	
		Change	Persistence
Simulation	Change	*Hits*	*False alarms*
	Persistence	*Misses*	*Correct rejections*
	Sum	*Prevalence*	1 – *Prevalence*

Figure 1 shows the layout of a contingency table with two categories: *Change* and *Persistence*. The terms *Hits*, *Misses*, *False Alarms* and *Correct Rejections* are used in this text. For clarification, they are defined in Table 1.

Definitions

Term	Definition
Hits	Cases where change is simulated correctly. Also known as true positives
Misses	Cases where reference change is simulated as persistence. Also known as false negatives
False Alarms	Cases where reference persistence is simulated as change. Also known as false positives
Correct Rejections	Cases where persistence is simulated correctly. Also known as true negatives

The above terms are defined as proportions, and sum to 1. In this case, the proportion of area of reference change (hits + misses) is called *Prevalence*. However, it is equally justifiable to define the terms of Table 1 as size. In that case, their sum is the total size of the study area, and the size of area of reference change (hits + misses) is referred to as *Abundance*. In this study, we use various tools of analysis. Our main and concluding analysis is better interpreted when the components of map comparison are described as proportions of the study area. Still, one of our other analyses uses a tool that calculates respective areas of those components. Describing each analysis, we clarify how it interprets components of map comparison. More detailed explanation about the contingency table can be found in several references, such as Pontius Jr and Parmentier (2014) in LUCC modelling literature, and Congalton (2004) in remote sensing literature.

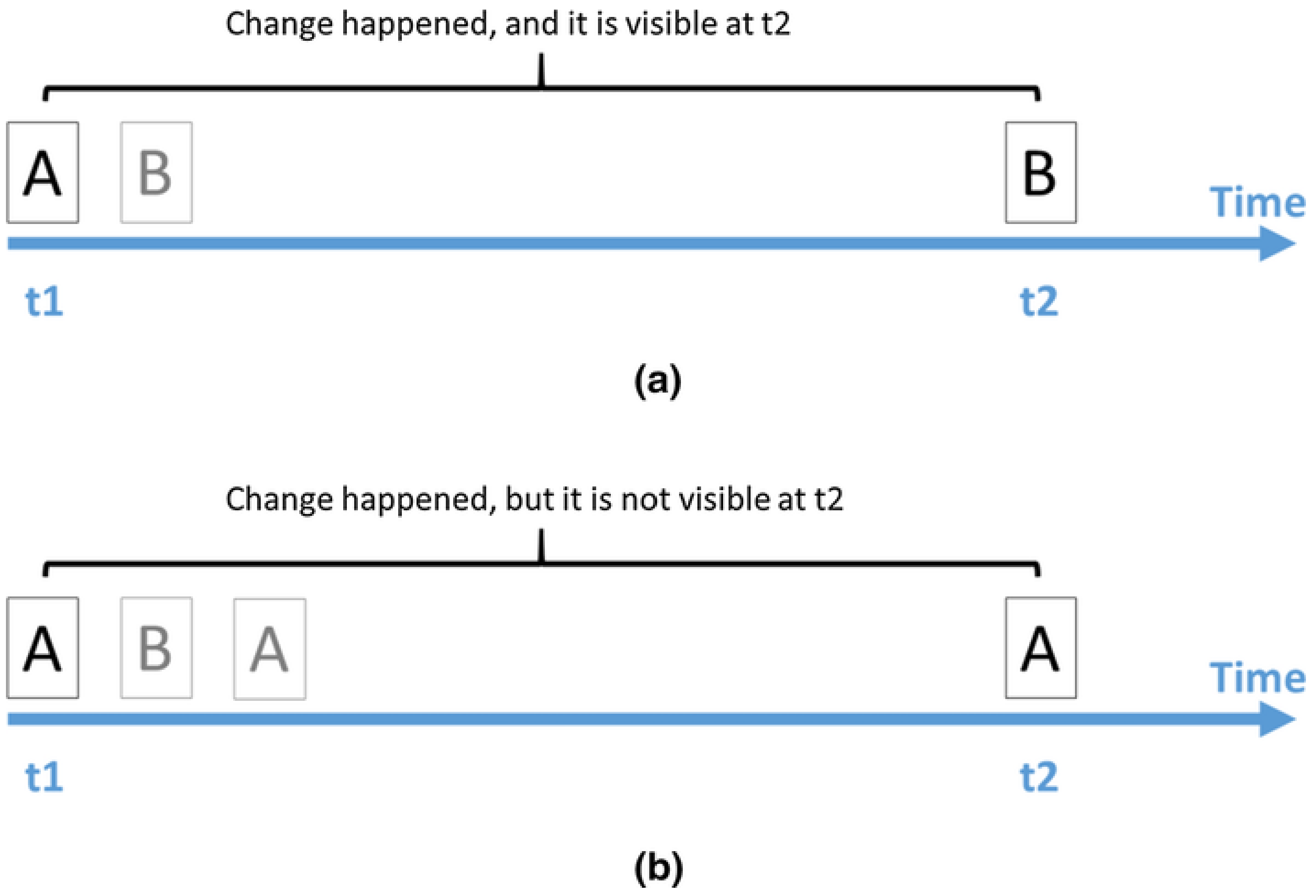
Scope and objectives

The scope of this paper is the validation of models of land change with a single transition and spatiotemporal dependency. Within this scope, we present a method for assessment of strengths and weaknesses of models in comparison with a baseline. In this section, we first describe the above-said scope. Then, we explain the reason for using baselines in model assessment. Next, we highlight challenges and problems of using baselines. Finally, we define the objective of the paper with respect to the above-said challenges and problems.

We begin by describing the scope. A single transition is a one-way change between two land classes. This means that the subset of land change models within the scope of this study includes binary models of processes of change that are one-way during the study time interval. Beyond the scope of this study, in models that involve gain of several classes, or in models that involve simultaneous gain and loss of the same class in different parts of the study area, correct prediction of change is more challenging (Pontius Jr et al. 2018), and model assessment is accordingly more complicated. Figure 2 demonstrates a hypothetical example, which shows an important difference between models within the scope of this study and models beyond the scope of this study. Figure 2 shows that for models beyond the scope of this study, even if a place appears unchanged, it may have undergone change several times. As such, assessment of models beyond the scope of this study involves an additional challenge in interpretation of maps of change. In contrast, the definition of scope of this study involves a simplifying assumption regarding land classes and transitions. Note that an implication of the assumption of one-way change is that the study area will be the region that was unchanged at the beginning of the time interval of validation of the model, as this region is the only candidate for change during the simulation.

Fig. 2

Implications of reversibility and irreversibility in interpretation of assessment data: **a** in a one-way process, a change of ~~state~~class during the study period can be identified by comparing reference data at the beginning and end of the ~~period~~time interval; **b** in a reversible process, change of ~~state~~class may happen during the study ~~period~~time interval without a trace in the reference data at the beginning and end of the ~~period~~time interval


Change happened, and it is visible at t2
A
B
B
Time
t1
t2
(a)
Change happened, but it is not visible at t2
A
B
A
A
Time
t1
t2
(b)

Another assumption that we make in defining the scope of this study is spatiotemporal dependency. Spatiotemporal dependency occurs when places influence and are influenced by their surrounding neighborhoods. This is an implication of the first law of geography (Tobler applications within the scope of this study are forest fires (Gaudreau et al. 2016), forest insect infestations (Perez et al. 2016), deforestation (Pontius Jr 2018), urban expansion (Pijanowski et al. 2005), and spread of contaminations (Di Gregorio et al. 1997). These examples, like many other phenomena in geography and landscape ecology, involve spatiotemporal dependency. Moreover, the process of change in each of these examples is one-way.

The combination of spatiotemporal dependency and one-way change has an implication about the above applications: it implies that these applications involve neighborhoods where change spreads. In the above applications, there are places which, at the beginning of the study time interval, can transmit the phenomenon to their surroundings. In other words, the phenomenon is present in those places at the beginning of the study. Throughout this paper, we refer to these places as the initial points of spread. For example, in applications such as forest fires or forest insect infestations, initial points of spread are places that underwent change immediately before the beginning of the study, such that at the start of the study, fire or insects were present in those places. In other applications such as deforestation, urban expansion, or soil contamination, initial points of spread are on the border between the unchanged and changed zones. In any case, the modeler can define the initial points of spread based on the nature of the studied phenomenon. We use the idea of initial points of spread in our methods.

Having described the scope of the study, we now explain the rationale and issues regarding the use of baselines in model assessment. It is desirable to identify strengths and weaknesses of models. However, strength and weakness are relative terms, and they make sense when they are expressed in a comparison of two or more things. It is a known approach of model assessment to compare the results of the model with another model as a baseline, also known as a benchmark (van Vliet et al. 2016). This comparison reveals whether the performance of the model is better or worse than the baseline. The use of baselines in assessment of LUCC models is not rare; in fact, in their review of 114 modeling works in LUCC literature, van Vliet et al. (2016) reported that 30% of the studied works involved baselines or benchmarks. Two questions need to be addressed before a baseline model can be used for assessment: (1) how to define the baseline model, and (2) how to compare the baseline model with the simulation. Regarding the first question, the modeler defines the baseline according to common sense and the particularities of the subject of study. For models within the scope of this paper, we define the baseline using the idea of neighborhoods around initial points of spread. We discuss this matter in detail in the methods section. Regarding the second question, some works in land change literature assess the baseline model in the same way that the simulation is assessed, and then compare the results of the two assessments (Pontius Jr et al. 2007; Pontius Jr 2018). This approach reveals useful information. For example, it answers questions such as: *Does the simulation make more Hits than the baseline? Which model makes more False Alarms? Which model makes more Misses?*

Although the above-said approach provides useful information for model assessment, it has a shortcoming, as it cannot answer more detailed questions such as: *How many of the Hits of the simulation are also Hits of the baseline? How many of the Hits of the simulation are missed by the baseline? How many of the Hits of the baseline are missed by the simulation?* Answering these questions reveals useful details about strengths and weaknesses of the model with respect to the baseline. For example, Hits of the simulation that are missed by the baseline show relative strength of the simulation with respect to the baseline. Moreover, for models within the scope of this paper, we want objective information about where the errors occur with respect to initial points of spread. To that end, we are interested in finding answers to questions such as: *How much of the model's correct predictions of change are near initial points of spread? What is the dominant type of error far from initial points of spread? Does the model underestimate or overestimate change near or far from initial points of spread?* These questions are not answered by existing methods of model assessment, and we address them in this paper.

Considering the above, the objectives of this study are: (1) to present a method for building a baseline model for applications within the scope of this study; and (2) to present a method for comparing a simulation with a baseline, in such detail that the comparison of the simulation and baseline provides answers to the questions raised in the previous paragraph. We demonstrate our methods by applying them to an example case of study, and we discuss the results that our methods reveal about the example case. We also apply existing methods in the validation of our example case, in order to compare the existing methods with our methods.

Case study

The motivation for this work comes from previous efforts to evaluate a model of forest insect infestations (Harati et al. 2020). The Mountain Pine Beetle (MPB) is a native wood-boring insect infesting forests of western North America. During the past two decades, MPB outbreaks have become epidemic and damaged over half of the commercial pine volume in the province of British Columbia (BC), Canada (Natural Resources Canada 2019). Each summer, the insects fly in search of new hosts to infest. In the earlier years of the outbreaks, the eastward spread of infestations was limited by the Rocky Mountains, but eventually MPB crossed this natural barrier and arrived in the neighboring province of Alberta (Natural Resources Canada 2019). Contrary to infestation, which kills trees, are processes of forest regrowth and succession. However, these processes are much slower than the spread of insect infestations. As such, in our case of modeling MPB outbreaks in a 6-year ~~period~~time interval, we assume that the processes of succession did not have enough time to make any change in infested forests. For this reason, we consider infestation a one-way process of land change.

In previous studies, we developed a model to simulate the spread of the insect in BC (Harati et al. 2020). The predictor variables of the MPB model were elevation, aspect, slope, surface ruggedness, and sums of surrounding infestations weighted by four different distance functions, calculated for infestation data of the year of start of simulation and its preceding year. We used various algorithms and calibration settings to train the model, and used it to predict the spread of infestations from 2008 to 2014, hence producing one simulation for each algorithm. The scope of the present paper is assessment and validation of two simulations produced by the MPB model, and their comparison with a third model that is defined as a baseline. Therefore, development of simulation algorithms and adjustment of their parameters are not within the scope of this paper. Rather, we discuss how we can compare simulation outputs with reference datasets in order to obtain information that is useful in model assessment.

Methods

Our model assessment approach in this study is to compare the simulation with a baseline. As suggested by Pontius Jr et al. (2007) This should be a hyperlinked reference, not plain text.

we define the baseline as a neighborhood in proximity of initial points of spread. Our baseline model predicts that areas near initial points of spread will change, and areas far from initial points of spread will not change. In this sense, the baseline model divides the study area into two strata with respect to proximity to initial points of spread. In the following subsections we describe how we build the baseline model, and how we compare it with the initial reference, final reference and simulation simultaneously. Through such simultaneous comparison of four maps, we obtain useful information for model assessment. In addition, to distinguish between our analysis and existing methods, we note highlights about the existing method of multiple resolution map comparison, which is used for analyzing allocation errors.

Distance analysis and proximity suitability map

The concept of our proximity suitability map is that places adjacent to initial points of spread are the most suitable candidates for future changes in the entire study area; in turn, places adjacent to those candidates are the next suitable ones; and so on. In this concept, suitability decreases as distance from initial points of spread increases. For ease of computation, we use Manhattan distances in making the proximity suitability map. By overlaying the reference dataset of the end of validation time interval on the map of Manhattan distances from initial points of spread, we extract information about percentage of reference infestations in various classes of distance from initial points of spread. This information provides a better insight ~~into the case study~~concerning various distances.

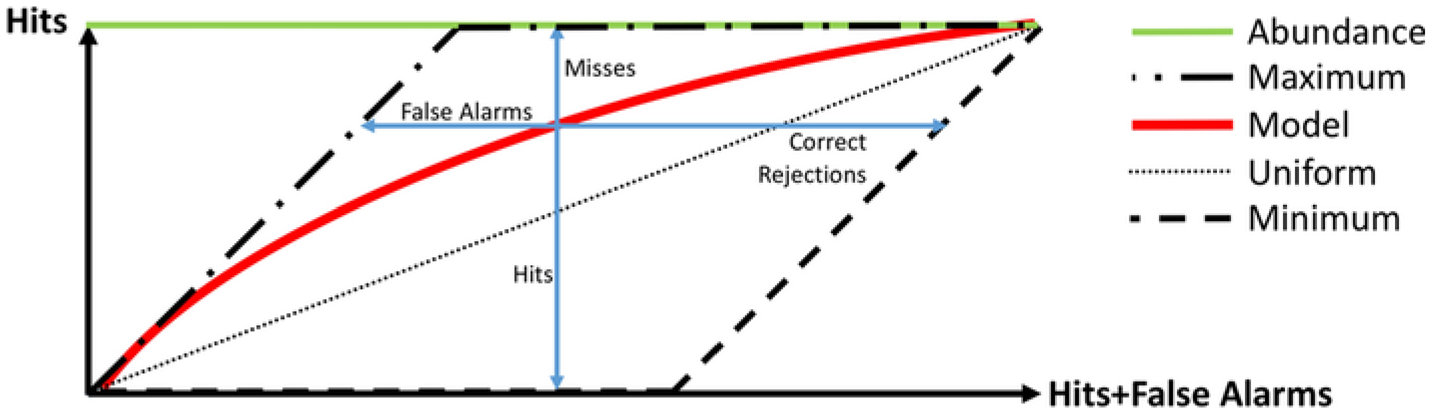
Total Operating Characteristic curves

We classify simulation and proximity suitability maps using thresholds. The classified predictions depend on their respective classification threshold. Such dependency is different for each model. We gain information about the performance of the models by analyzing their agreement and disagreement with reference data for each classification threshold. We perform and summarize this analysis using the Total Operating Characteristic (TOC) curve (Pontius Jr and Si 2014).

The TOC curve is a tool for analyzing agreement between the ranked output of a model with a binary reference. In this sense, its overall aim is similar to that of the Relative Operating Characteristic (ROC) curve. The TOC curve is obtained by connecting points that each represent a threshold for classification of model output. Corresponding to each threshold, there is a contingency table that summarizes agreement and disagreement between model output and reference data. Misses, hits, false alarms, and correct rejections for each classification are directly identified in the TOC curve as shown in Fig. 3. This is because of the way the plot coordinates and axes are defined. The TOC shows misses, hits, false alarms, and correct rejections as sizes; their sum equals the size of the study area. For each point of the TOC curve, the horizontal coordinate equals the sum of hits and false alarms of that point's respective contingency table, and the vertical coordinate equals hits. Similar to the ROC curve, the Area Under Curve (AUC) of the TOC curve offers a metric to summarise accuracy. In the TOC plot, a parallelogram bounds all possible TOC curves, as shown in Fig. 3. In this plot, the AUC is the ratio of the area under TOC curve in the parallelogram to the total area of the parallelogram.

Fig. 3

A hypothetical TOC curve. In this plot, hits, misses, false alarms, and correct rejections are measures of area. Abundance is the area of reference change



To find the threshold for each model, we identify the intersection of its TOC curve with the vertical line, $x = Abundance$, which is drawn down from the upper left corner of the parallelogram. On that vertical line, misses and false alarms are equal. Recalling that the sum of misses and hits is the quantity of reference change, and that the sum of false alarms and hits is the quantity of simulated change, it follows that at this intersection point, the quantity of change from the start to the end of the validation time interval equals the quantity of change in the reference data.

The distance threshold that we find for the proximity suitability map using the TOC curve is the key to define near and far with respect to initial points of spread. This threshold classifies the proximity suitability map into a binary map. That is, in the proximity suitability map, values less than or equal to that threshold are classified as *changed*, and other values are classified as *unchanged*. We use this classified map as the baseline model.

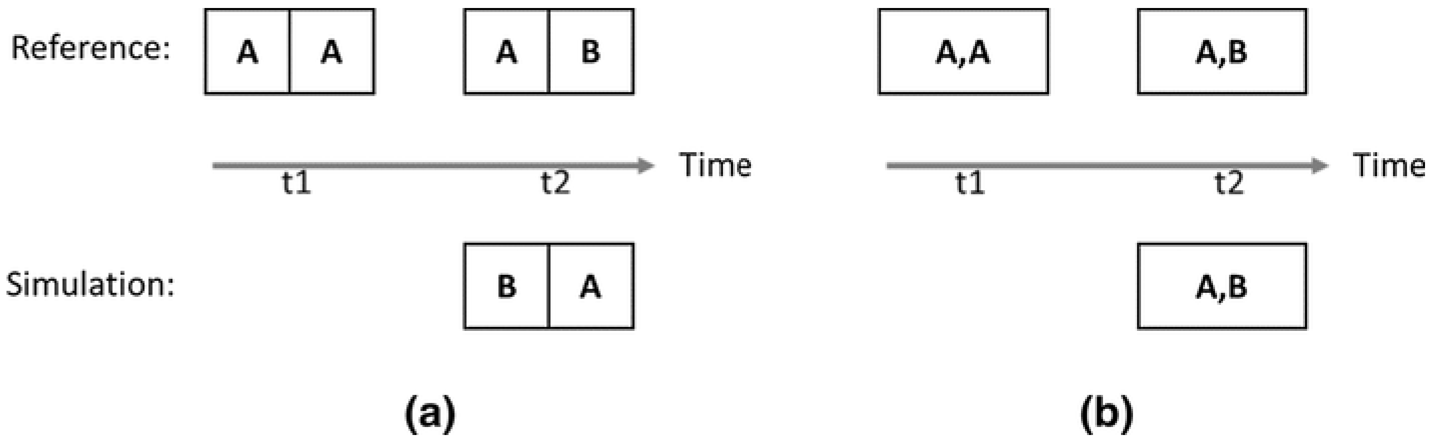
Three-map comparison at multiple resolutions

In order to assess performance of the models, existing methods compare simultaneously three maps: reference at the beginning of the validation time interval, reference at the end of the validation time interval, and simulation at the end of the validation time interval. Such three-map comparison, which is well described in literature (Pontius Jr et al. 2004, 2008), is based on the contingency table of simulated change versus reference change. In applications with change that is one-way during the validation time interval, places that underwent change before the beginning of the study are excluded from the study area, as they are no longer candidates for change.

By ~~classifying~~selecting the threshold at the correct quantity of change in suitability maps, we obtain binary maps in which quantification error is eliminated or minimized. The error that remains in model outputs is of allocation type. This implies that the size of misses equals interested in knowing the distances between such pairs of miss and false alarm errors. In comparison of two models whose quantification error is eliminated, the one whose allocation errors occur in shorter distances performs better. In order to include this consideration in model assessment, we use the method of three-map comparison at multiple resolutions (Pontius Jr 2002; Pontius Jr et al. 2004, 2008, 2011). In this method, fine-resolution maps are coarsened to a given resolution; then, they are compared. Throughout coarsening, several fine pixels are aggregated in a coarse pixel. In the coarse pixel, information concerning quantities remain, but information concerning allocation is eliminated. This effect is shown in Fig. 4. If two fine pixels containing a miss and a false alarm are put together in a coarse pixel, coarse resolution comparisons shows that both the coarsened reference and the coarsened simulation contain one unit of change and one unit of persistence in the aggregated pixel. As such, through coarsening, for each pair of miss and false alarm errors aggregated in a larger pixel, the counts of misses and false alarms each decrease by one, and the counts of hits and correct rejections each increase by one.

Fig. 4

A hypothetical example showing reduction of allocation error in coarsening of resolution: (a) two fine-resolution pixels contain a miss and a false alarm; (b) when aggregated in the same coarse pixel, they indicate a hit and a correct rejection. ‘A’ and ‘B’ are ~~states~~classes



Four-map comparison and analysis of components of change in partitioned study area

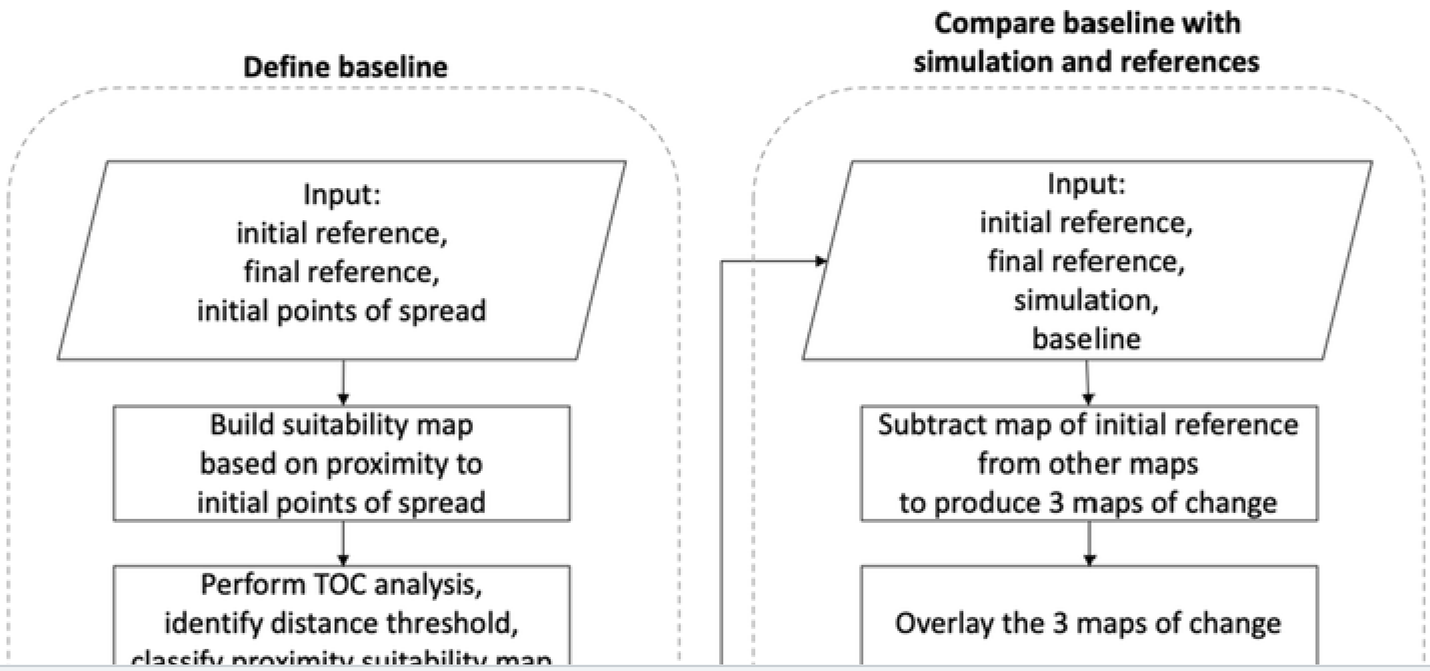
The baseline model divides the study area into two strata: one that is closer to initial points of spread, and another that is ~~further~~farther from initial points of spread. For ease of reference, we call them the *near* and *far* strata, respectively. The baseline model predicts the near stratum as changed, and the far stratum as unchanged. These strata are subsets of the study area, such that their union is the entire study area and their intersection is empty. This stratification is important for us because the near stratum is more suitable than the far stratum for future change, and we expect that the simulation shows this ~~matter~~pattern. To assess the model using baseline, we simultaneously compare the four maps of initial reference, final reference, simulation, and baseline. We call this procedure the four-map comparison.

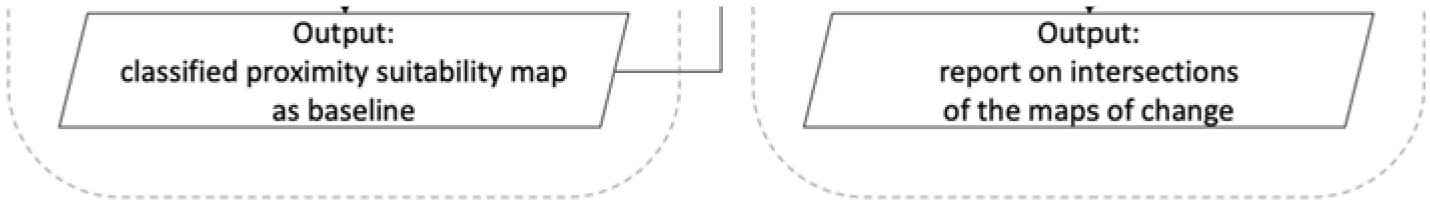
The four-map comparison is performed by first calculating the difference between the map of reference data at the beginning of the study and each of the three other maps. After this step, there will be three maps of change from beginning to end of the validation time interval. The three maps of change are: reference, simulation, and baseline. Next, we note that each of these three maps is a set of pixels, and we overlap the three sets and note their unions and intersections. Since the scope of our study is limited to models with two ~~states~~classes, the combination of these three sets can produce up to 8 subsets. For example, the subset of pixels that indicate change in the reference set, change in the simulation set, and change in the baseline set ~~include~~sis Hits in proximity to the initial points of spread. Other subsets are interpreted similarly. Of the 8 subsets thus produced, 2 of them include no change in the reference set and no change in the simulation set. These two are the subsets of *Correct Rejections*, and we do not include them in the rest of our analysis, as they do not provide any information about correct or incorrect simulations involving change. The remaining 6 subsets indicate respective areas of *Misses*, *Hits*, and *False alarms* in the *near* stratum and in the *far* stratum.

Our validation method involves (1) defining a proximity baseline and (2) comparing the baseline with simulation and references. A summary of steps of these two activities is depicted in Fig. 5. Note that for applications where scientists choose a different baseline instead of the proximity baseline, the steps for comparison of the baseline with simulation and references are the same as those shown in the figure.

Fig. 5

Flowchart of the proposed validation method including definition of a proximity baseline and use of the baseline in assessment of simulations





Data

We assess a MPB model (Perez et al. 2016) and compare two simulations generated by different algorithms, namely, logistic regression (LR) and random forest (RF). The simulations predict the spread of MPB infestations in the forests of BC from 2008 to 2014. Each simulation is analysed in comparison with reference data of the years 2008 and 2014. In addition to the two aforementioned simulations, we define a baseline model and compare its output with reference datasets and LR and RF simulations.

All model outputs and reference datasets are rasters of the study area. Dimensions of the maps are 4011 rows by 3516 columns ~~by 4011 pixels~~, with pixel resolution of 400 m. The extent of the study area is from 59°59′27″ N 138°54′19″ W to 48°59′53″ N 114°2′37″ W. The study area is the union of forest areas that were not infested at the beginning of 2008. In other words, areas that were already infested before 2008 were excluded from analysis because no change would happen in ~~the state of~~ those areas.

Reference datasets are binary, meaning each pixel of reference data is classified as *infested* or *not infested* at 2008 or 2014. The two MPB model outputs are suitability maps for the region that is not infested at 2008. For two pixels in the same suitability map, the one with the higher value is a more suitable candidate for infestations beyond 2008. The value of a pixel in a suitability map is comparable in terms of rank with values of other pixels in the same map. It is meaningless to compare the value of a pixel in a suitability map with the value of the corresponding pixel in the other suitability map because the suitability values are not probabilities.

The baseline model, which is also associated with a suitability map, is built as described in the Methods section. The input required for this model is the initial points of spread, i.e. places of insects in reference data of 2008, which is also a raster with the same dimensions and resolution as the other datasets used in this study.

Reference infestation data was obtained from the web portal of BC Ministry of Forests, Lands, Natural Resource Operations and Rural Development—hereinafter The Ministry (Province of British Columbia 2015). The data was then converted to binary using a threshold. The Ministry maintains datasets of cumulative MPB attacks based on the Provincial Aerial Overview of Forest Health (Province of British Columbia 2020). Detection of areas of MPB infestation is based on the fact that patches of trees attacked by the insect, change color in subsequent years.

Software tools

Our analyses were carried out using packages “raster” (Hijmans 2019), “lulcc” (Moulds et al. 2015), and “TOC” (Pontius Jr et al. 2015) of the statistical software R (R Core Team 2019). Maps were produced with ArcGIS for Desktop (ESRI 2015). TOC output plot was produced using TOC Generator software package (Liu 2020).

Results

This section presents the results of applying the methods of the paper on the example case of MPB infestations.

Table 2 shows the distribution of reference changes in various classes of Manhattan distance from the initial points of spread. The median Manhattan distance to nearest initial point of spread was 3 pixels. This distance corresponds to the 6-year duration, i.e. from 2008 to 2014. This table includes two noteworthy findings: firstly, in the 6 years of validation time interval 75% of the new infestations occurred within only 8 pixels (Manhattan distance) from initial points of spread; and secondly, the remaining 25% of new infestations were dispersed as far as 876 pixels from initial points of spread.

Table 2

Percentiles of shortest Manhattan distances to initial points of spread

Manhattan distance (pixels)	Percentile (%)
1	0~~.0~~
2	25~~.0~~
3	50~~.0~~
8	75~~.0~~
876	100~~.0~~

TOC curves for two simulations as well as the baseline model are shown in Fig. 6. It can be seen that the baseline model has a higher AUC than the other two models, and the RF model has a higher AUC than the LR model. At about 2/3 of the maximum height of the plot area, the TOC curve of the LR model includes a bend that is sharper than the RF curve. To the right of that bend in LR TOC plot, i.e., with lower suitability thresholds that classify more pixels as infested, changes in threshold cause larger changes in false alarms and smaller changes in hits compared to the other two models. Using these curves, we also selected the classification thresholds for each of the suitability maps. Each of these thresholds corresponds to the point on each curve that is under the upper-left corner of the TOC plot parallelogram. In particular, in our TOC curve analysis, the identified distance threshold for classification of the proximity suitability map was 2 pixels.

Fig. 6

are 0.79 for Logistic Regression, 0.85 for Random Forest, and 0.86 for Proximity

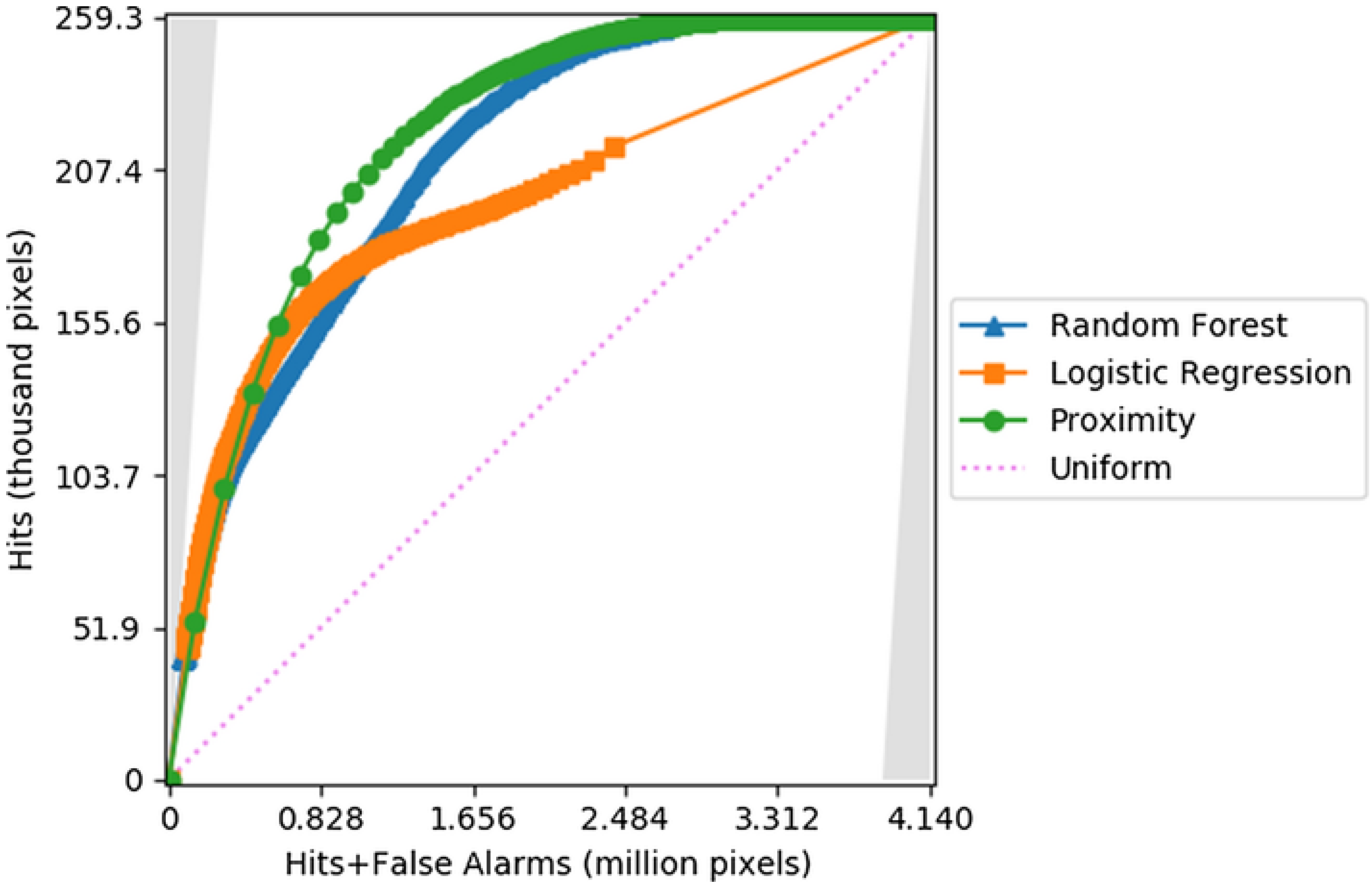


Figure 7 shows the proximity suitability map, which was calculated based on Manhattan distances to initial points of spread. In addition, this figure shows the RF and LR suitability maps, which were the input data of our analysis. The RF and LR maps were produced in a previous study, based on infestation and geographical data (Harati et al. 2020). The maps in this figure are not classified, and each of them indicates some parts of the study area as more suitable for infestations than other parts of the study area. We used the thresholds obtained in TOC analysis to classify the maps of Fig. 7. The results of these classifications were binary maps in which each data pixel is infested or not infested. Figure 8 shows the components of agreement and disagreement with reference data for the classified RF, classified LR, and baseline. The baseline map was produced by classifying the proximity suitability map using a threshold of 2 pixels. Each of the maps of Fig. 8 is the result of comparison of maps of reference at 2008, reference at 2014, and the respective model at 2014. The figure shows that most of the observed and simulated change happened in the northern part of the study area. The figure also shows that all models missed changes in the northeast of the study area.

Fig. 7

Maps of infestation suitability from 2008 to 2014 for two simulations and a proximity model. Initial points of outbreak are where insects were at start (2008). Analyses exclude infestations prior to 2008

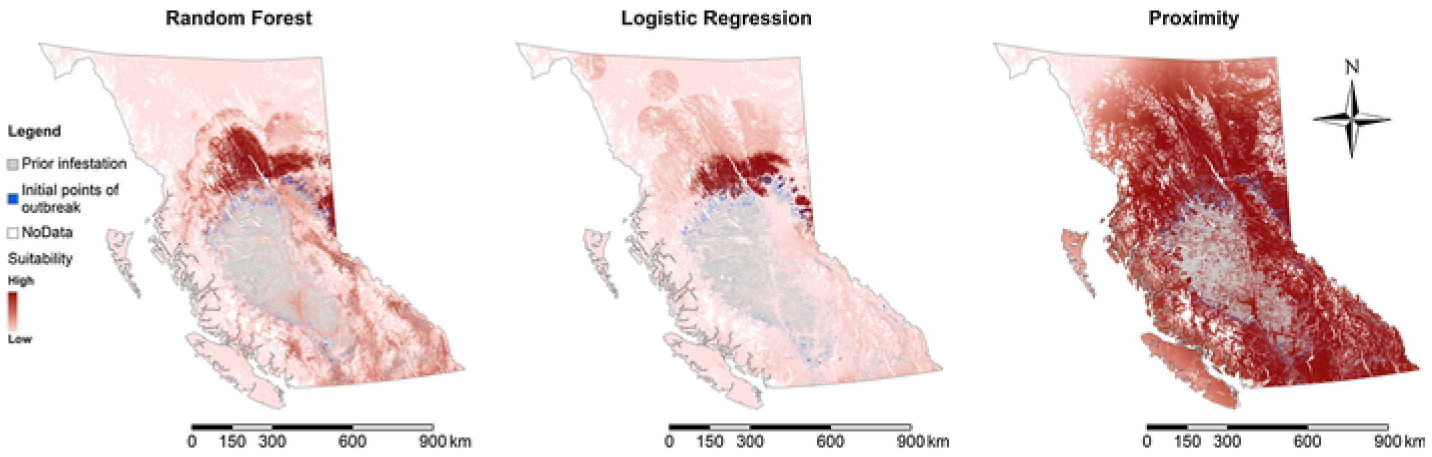


Fig. 8

Comparison of maps of reference in 2008, reference in 2014, and prediction in 2014 for two simulations and a baseline model. Initial points of outbreak are where insects were at start (2008). Analyses exclude infestations prior to 2008

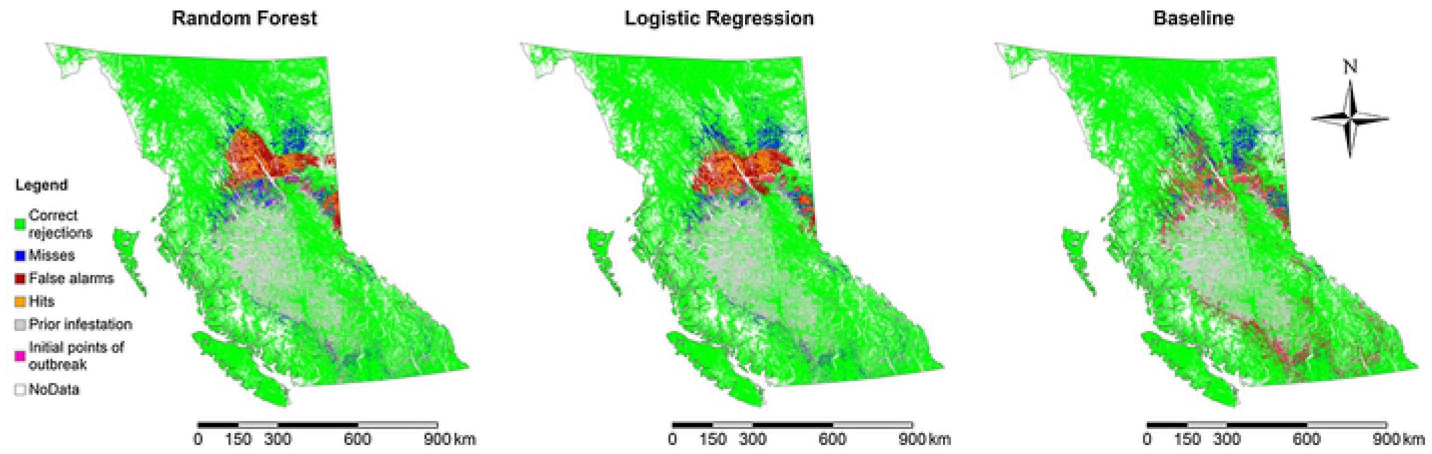
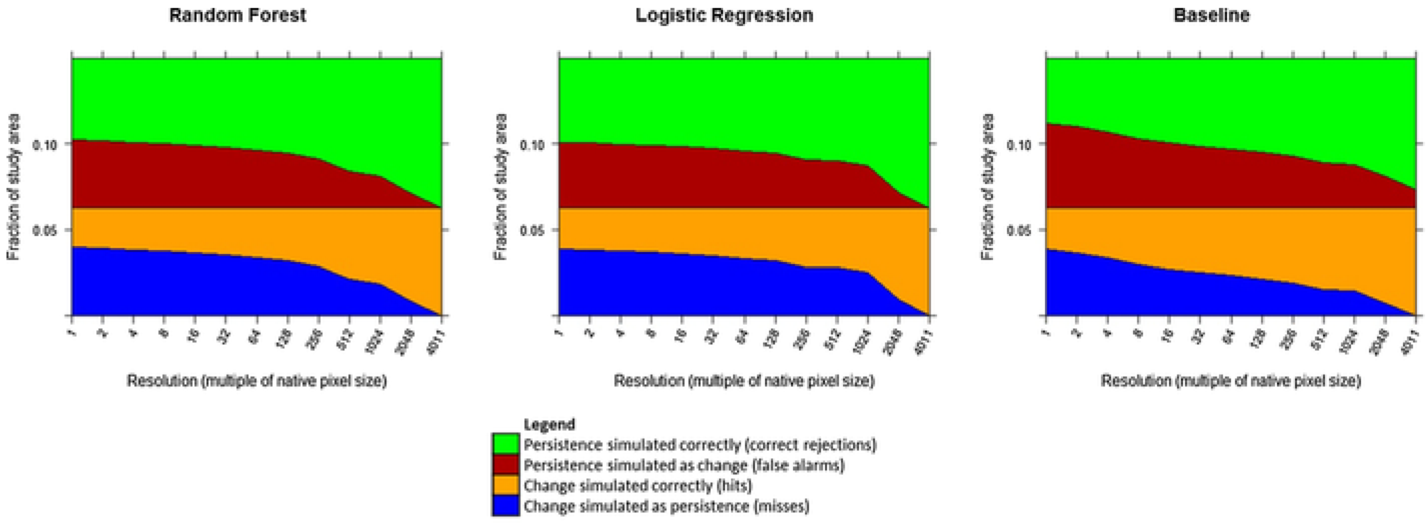


Figure 9 summarizes the results of three-map comparisons at multiple resolutions for each model. This figure shows area proportions of hits, misses, false alarms, and correct rejections of each simulation at multiple resolutions. The figure shows that the reference infestation is are zero at the coarsest resolution. Baseline has more false alarms than misses, which indicates that the baseline model has more simulated infestation than reference infestation.

Fig. 9

Multiple-resolution analysis of components of agreement and disagreement between maps of reference in 2008, reference in 2014, and prediction in 2014



In Fig. 10 we demonstrate the result of partitioning of the study area using the baseline model. The near and far strata are defined by the threshold obtained in the analysis of the TOC curve of the proximity suitability map, which was 2 pixels. The baseline model predicts the near and far strata as infested and not infested, respectively.

Fig. 10

Partition of the study area into *near* and *far* strata using the baseline model. Initial points of outbreak are where insects were at start (2008). Analyses exclude infestations prior to 2008

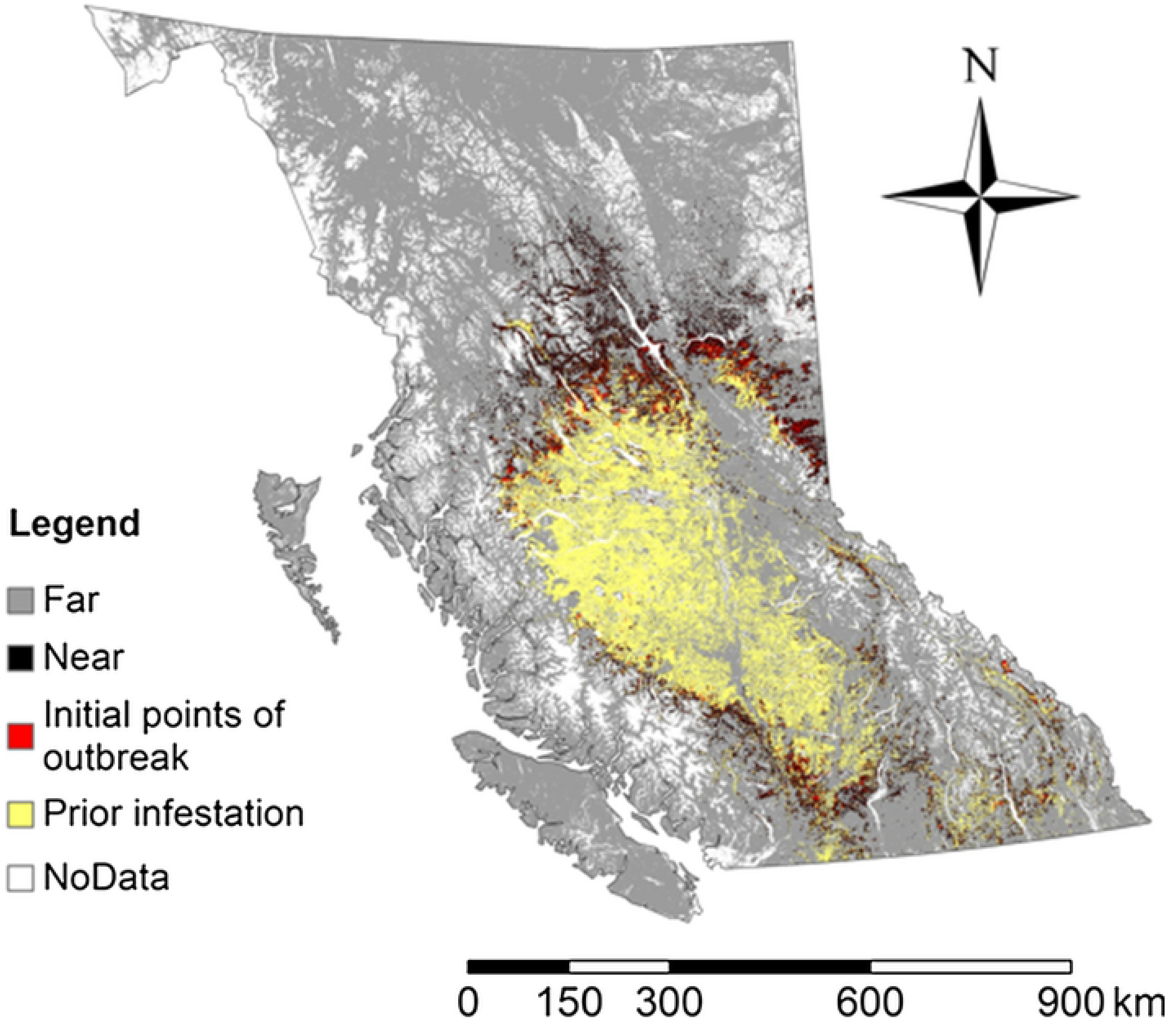
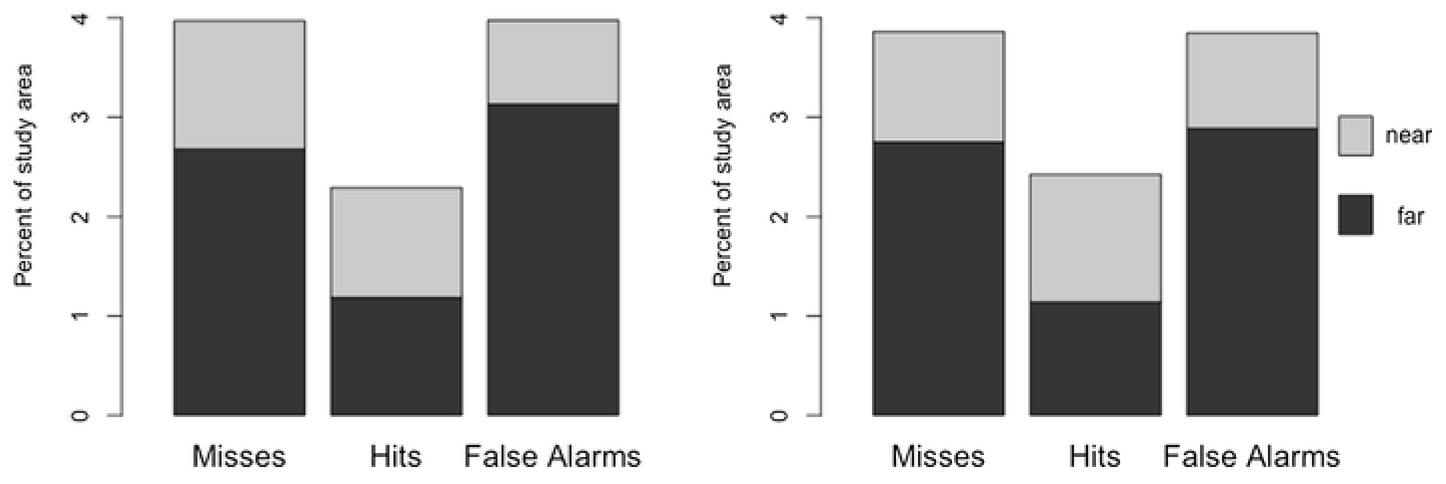


Figure 11 shows the components of agreement and disagreement for simulated and reference change in the two strata of the study area that are defined based on proximity to initial points of spread. These strata are the predictions of the baseline model. The misses, hits, and false alarms of each model in near stratum are areas that the baseline model predicts as infested. Conversely, misses, hits, and false alarms of each model in the far stratum are areas that the baseline model predicts as not infested. RF has ~~more allocation error~~fewer hits than LR. The overall sizes of misses and false alarms of RF are equal. Similarly, the overall sizes of misses and false alarms of LR are equal. In the near stratum of the study area, RF has more misses than false alarms.

Fig. 11

Components of change in near and far strata of the study area



Discussion

This section describes the information and insight from the results for the example case of MPB infestations. These descriptions also include results of existing methods to demonstrate how the new methods complement existing methods by providing information that would otherwise not be attainable. Afterwards, we discuss the limitations and implications of the methods developed in this study.

Insights about the case of study

The results of distance analysis in Table 2 reveal interesting information about reference data. The fact that the median distance of spread in 6 years is 3 pixels shows that many new infestations occurred near initial points of spread. This observation suggests it is reasonable to build the baseline in proximity to initial points of spread. On a different note, it is also worth mentioning that the Manhattan distance to initial points of spread ranged from 1 to 876 pixels. The upper end of this range is much larger than the lower 75 percent of the values, which raises questions about the possible cause of such difference. Future studies can analyse this case and assess if such large distance of MPB spread has been due to, for example, weather and wind conditions, man-made interventions (e.g. MPB piggy-backing on logs transported to processing plants), or data processing errors.

In the analysis of the TOC curves, it is notable that the proximity suitability map has the highest AUC, which indicates better performance compared to the other two models. The curves also show that LR is above RF at the correct quantity, which indicates that LR is better than RF at specifying where ~~few~~ infestations exist. LR curve is below the others in the right side of the TOC space, which indicates that the non-LR models are better at assigning low ranking suitability values to pixels where infestation does not occur.

The distance threshold obtained in the analysis of the TOC curve of the proximity suitability map defines the near and far strata of the study area. In this case, the areas within Manhattan distance of 2 pixels from an initial point of spread are considered the near stratum. Conversely, the areas with Manhattan distance of more than 2 pixels from an initial point of spread are considered the far stratum. In this way, the baseline model is built. The baseline model predicts change in all of the near stratum, and no change in the far stratum. In other words, the baseline model predicts that the near stratum will be completely infested, and the far stratum will not be infested.

Visual inspection of Fig. 8 shows RF and LR simulations missed many changes in the region in the northeast of the study area. This region is located on the eastern side of the Rocky Mountains. The Rocky Mountains previously served as a natural barrier blocking the spread of MPB towards the east. However, during the validation time interval, large MPB outbreaks occurred in the northeastern part of the province. We see in Fig. 8 that the RF and LR models did not predict these outbreaks.

Figure 9 shows that RF and LR models have no quantity error. This is because in the analysis of their TOC curves, classification thresholds were selected to eliminate quantity error. The figure shows that for the baseline model, the elimination of errors at the coarsest resolution does not happen completely. The baseline model has some excess false alarms, which means that its quantity error was not completely removed. This is because the proximity suitability is constructed using Manhattan distances from initial points of spread. Manhattan distances are expressed as natural numbers, and there are a large number of pixels having the same distance value. Nevertheless, the selection of the threshold has been set such that the quantification error of the classified result was minimized. That is, the quantification error for the selected threshold is smaller than the quantification error for any other threshold.

In Fig. 9, as resolutions become coarser, misses and false alarms decrease, and hits and correct rejections increase. A closer look at the LR plot in Fig. 9 shows that in the coarsening from cell-size 128 to cell-size 256 the curves of errors are steeper than in smaller cell-sizes. This shows that at this particular resolution, there are suddenly more pairs of miss and false alarm errors that cancel one another. In the classified LR simulation, the pairs of miss and false alarm errors that are located between 128 and 256 pixels away from one another are more noticeable than those in shorter distances from one another. This distance interval is an indicator of allocation error of the LR simulation. Following similar steps with the RF simulation, we find that its respective distance interval is between 256 and 512 pixels. This means that, compared with the LR simulation, allocation errors in the RF simulation are further away from one another. In other words, multiple-resolution analysis revealed that the LR model performs better than the RF model in terms of allocation error for that particular coarsening of resolution.

The four-map comparison provides information that the existing methods do not show. Recall that the four-map comparison includes reviewing the results of three-map comparison in the two strata defined by the baseline model (~~near~~ near and ~~far~~ far strata with respect to initial points of spread, as shown in Fig. 10). This method reveals new information about performance of the models. Figure 11 shows the components of change for each model in the study area, as well as their breakdown in the two strata. For each simulation, this figure summarizes useful information obtained from comparing four maps: reference 2008, reference 2014, respective simulation, and the baseline model. The figure provides information on where the ~~misses~~Misses, ~~hits~~Hits, and ~~false alarms~~False Alarms are in the map; how much of them is in the ~~*near*~~ near stratum, and how much in the ~~*far*~~ far stratum. Note that the sum of each ~~column~~bar in Fig. 11 corresponds with the fine-resolution information given by Fig. 9. However, the breakdown of each ~~column~~bar into ~~*near*~~ near and ~~*far*~~ far strata includes new information that cannot be found from Fig. 9. For one thing, it shows how each of the simulations compares with the baseline model. Since the baseline model predicts no change in the ~~*far*~~ far stratum, it misses all reference changes in the ~~*far*~~ far stratum. As such, the ~~*Misses*~~ Misses of simulations in the ~~*far*~~ far stratum are ~~*Misses*~~ Misses of the baseline model as well. ~~*Hits*~~ Hits in the ~~*far*~~ far stratum show the strength of the LR and RF simulations over the baseline model. On the other hand, LR and RF simulations also produce ~~false alarms~~False Alarms in the ~~*far*~~ far stratum, which the baseline model does not. Similar arguments can be stated about the near ~~*near*~~ stratum. The baseline model predicts change in all of the ~~*near*~~ near stratum, therefore ~~it~~the baseline model misses nothing there. ~~*Misses*~~ Misses of the simulation in the ~~*near*~~ near stratum are all ~~*Hits*~~ Hits of the baseline model, and indicate relative weakness of the simulation with respect to the baseline model. On the other hand, the baseline model produces more ~~false alarms~~False Alarms than the simulation in the ~~*near*~~ near stratum.

In the example case of MPB infestations, in addition to comparing the simulations with the baseline model, we can gain insight about the models by noting how they compare against one another. ~~Because both~~Both LR and RF simulations are classified such that their quantification errors are almost eliminated~~;~~ for each simulation, meaning misses and false alarms in the entire study area are nearly equal. However, the distribution of these errors in the two strata of the study area is noteworthy. In both simulations, in the near stratum there are more misses than false alarms; and in the far stratum there are fewer misses than false alarms. This difference means that in both simulations, in the near stratum the quantity of change is predicted less than in reference data; and in the far stratum the quantity of change is predicted more than in reference data. In other words, both simulations involve errors of allocating less change than reference to the near stratum, and allocating more change than reference to the far stratum. Moreover, this error is larger in the RF simulation than in the LR simulation. The RF model, in comparison with the LR model, has the weakness of underestimating the spread of infestations near initial points of spread. This useful finding was not evident in previous analyses; rather, it is the result of the additional analysis of components of change in near and far strata.

Our analysis of allocation errors described in this section depends on the strata defined by the baseline model, which, in turn, are the result of distance analysis. As such, for a different result of distance analysis from what we calculated, our assessment of the simulations could be different from what we presented in this section.

Limitations and implications of the study

In this paper we developed methods to assess land change simulation with respect to the baseline that is constructed with common sense. Our motivation in this study was to give an objective answer to the question of where errors occur. A map of errors, of course, shows where errors occur. However, the interpretation of maps is subjective (van Vliet et al. 2011). We sought methods to extract objective information from data that was available in initial reference, final reference, simulation, and baseline maps. Our methods have two ~~implications~~new concepts for assessment of land change models: (1) use of near and far strata as defined in this paper for models within the scope of this study; and (2) simultaneous comparison of the four maps of initial reference, final reference, simulation, and baseline, for models that are assessed with a baseline.

This paper demonstrated the application of our methods on an example case of forest insect infestations, but the implications of this study are not limited to this case. The methods of this study can be applied in assessment of a variety of models involving a single transition and spatiotemporal dependency. Nevertheless, the scope of this paper is one of its limits, as there are a wide range of applications that involve multiple land classes and reversible transitions, creating more complicated assessment problems. These problems deserve to be addressed in future works.

An important point to consider about the methods proposed in this paper is that even within the scope of the study, distance from initial points of spread is not the only factor that may be related to errors. In particular, in phenomena with factors that act along certain vectors, it is possible to observe patterns that defy the assumption of spread of change in proximity of previously changed places. Examples of these are the effect of wind on ~~of~~ forest fires and forest insect infestations, and the effect of roads on expansion of cities. Fire and insect infestations can spread rapidly along wind vectors, and urban built areas can spread along roads at a faster pace than what a baseline model might suggest. Modelers should be mindful of such effects when interpreting results of model assessment methods. It is important to consider, though, that in such cases if the modelers have a reasonable idea of the other factors that influence the spread of change, they can include that idea in the construction of the baseline model, so that the baseline model agrees with common sense and is still easy to understand. Then ~~they~~modelers can use the method of four-map comparison to assess their simulation using the new baseline. In all cases, the rationale of this paper is to use appropriate baseline models in order to gain insights concerning the performance of simulations. If the process of change is known to happen along certain vectors, then the baseline model can be constructed as predicting change along those vectors. If the process of change is known to include a combination of vector effects and proximity effects, then the baseline model can be constructed by superposition of a vector baseline and a proximity baseline.

Another matter worth mentioning is that the use of Manhattan distance involves some deviation with respect to Euclidean distance. The calculation of Manhattan distance has a computational advantage over Euclidean distance in applications where maps have a large number of rows and columns. However, with modern-day hardware and software, calculation of Euclidean distance is feasible for many applications. Modelers should be mindful of this matter in their assessments. In our case, even though we used Manhattan distance to build our proximity suitability map, the ~~respective~~proximity TOC curve had the highest AUC of all models. This curve was remarkably higher than the random line, which shows that our baseline was much more accurate than a random baseline. The curve was also higher than the RF curve at all threshold points. We built our baseline as an easy-to-understand model that makes sense, especially more relevant than a random model, because we know that infestations do not spread randomly. Use of random baselines has been criticized as irrelevant and/or misleading (Pontius Jr and Millones 2011) This citation should include a hyperlink to the reference, which is already in the list of references. This citation should not be plain text. Please convert the text into a hyperlink referring to the respective reference. . The TOC curves indicate that our baseline in this application offered helpful insight. Nevertheless, the building of our baseline involved simplifications, and with simplifications come inaccuracies.

Presently, the method of three-map comparison at multiple resolutions serves this purpose. However, analysis done by the multiple resolution method is dependent on the relative coordinates of pixels with respect to a corner of the map of the study area. This means that the result of multiple resolution analysis of a map can change especially at coarser resolutions, depending on the reference point for defining coordinates of map pixels. Our ~~proposed~~strata method in this paper does not depend on a fixed point on the map. Rather, our analysis is based on the phenomenon under study. This can be the basis for development of new methods for assessment of allocation error, while avoiding the said problem with multiple resolution analysis. Such methods could be developed by expanding our method to partition the study area into multiple strata instead of two.

Conclusions

Existing methods for evaluation of allocation error provide information on how far allocation errors are with respect to one another, but not with respect to ~~reference data~~a baseline model. Our article addressed the topic of allocation of errors in simulations of a subset of land change processes with a single transition and neighborhood effects. For applications in this subset, we presented methods to obtain objective information about where errors occur, by partitioning the study area with a baseline model that is defined through distance analysis of reference data, and performing a four-map comparison including the baseline model, the simulation, the initial reference, and the final reference. These methods identify the distribution of misses, hits, and false alarms in the two strata of the study area, which are defined based on proximity to initial points of spread of the phenomenon. The methods of this paper helped us gain insight concerning the performance of two example simulation cases, and revealed information that would otherwise be unattainable. We recommend that for a one-way, geographically-spreading process, model validation should distinguish between errors that are allocated near and far from the initial points of spread.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements

We are thankful to the Natural Sciences and Engineering Research Council (NSERC) of Canada for partial support of this study under the Discovery Grant Program awarded to LP, and to the Université de Montréal's International Affairs Office (IAO) for their financial support through the International Partnership Development program, which allowed the collaboration between researchers from UdeM and CREAF. RMH received financial support from the NEWFOREST (PIRSES-GA-2013-612645) program of the European Union's Seventh Framework Programme. The United States National Science Foundation supported RGP through its Long Term Ecological Research Network via grant OCE-1637630 for Plum Island Ecosystems.

Data availability

The datasets generated and/or analysed are available in the Open Science Framework repository, via https://osf.io/d5em3/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Batty M, Torrens PM (2005) Modelling and prediction in a complex world. Futures 37:745–766

Brown DG, Verburg PH, Pontius RG Jr, Lange MD (2013) Opportunities to improve impact, integration, and evaluation of land change models. Curr Opin Environ Sustain 5:452–457

Brown DG, Walker R, Manson S, Seto K (2012) Modeling land use and land cover change. In: Gutman G, Janetos AC, Justice CO et al (eds) Land change science. Springer, Dordrecht, pp 395–409

Chen H, Pontius RG Jr (2010) Diagnostic tools to evaluate a spatial land change projection along a gradient of an explanatory variable. Landsc Ecol 25:1319–1331

Congalton RG (2004) Putting the map back in map accuracy assessment. In: Lunetta RS, Lyon JG (eds) Remote sensing and GIS accuracy assessment. CRC Press, Boca Raton, pp 1–11

Cushman SA, Macdonald EA, Landguth EL et al (2017) Multiple-scale prediction of forest loss risk across Borneo. Landsc Ecol 32:1581–1598

de Sousa-Neto ER, Gomes L, Nascimento N et al (2018) Land use and land cover transition in Brazil and their effects on greenhouse gas emissions. Soil management and climate change. Academic Press, Cambridge, pp 309–321

Di Gregorio S, Serra R, Villani M (1997) A cellular automata model of soil bioremediation. Complex Syst 11:31–54

ESRI (2015) ArcGIS 10.4.1 for desktop

Foody GM (2004) Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. Photogramm

Gaudreau J, Perez L, Drapeau P (2016) BorealFireSim: a GIS-based cellular automata model of wildfires for the boreal forest of Quebec in a climate change paradigm. Ecol Inform 32:12–27

Hagen-Zanker A (2006) Map comparison methods that simultaneously address overlap and structure. J Geogr Syst 8:165–185

Harati S, Perez L, Molowny-Horas R (2020) Integrating neighborhood effect and supervised machine learning techniques to model and simulate forest insect outbreaks in british columbia, canada. Forests 11:1–23

Hermoso V, Morán-Ordóñez A, Brotons L (2018) Assessing the role of Natura 2000 at maintaining dynamic landscapes in Europe over the last two decades: implications for conservation. Landsc Ecol 33:1447–1460

Hijmans RJ (2019) raster: geographic data analysis and modeling. R package version 2.9–5

Lambin EF, Geist H, Rindfuss RR (2006) Introduction: local processes with global impacts. In: Lambin EF, Geist H (eds) Land-use and land-cover change. Springer, Berlin, pp 1–8

Li Z, Huffman T, Zhang A et al (2012) Spatially locating soil classes within complex soil polygons – Mapping soil capability for agriculture in Saskatchewan Canada. Agric Ecosyst Environ 152:59–67

Liu Z (2020) TOC Curve Generator. https://lazygis.github.io/projects/TOCCurveGenerator

Moulds S, Buytaert W, Mijic A (2015) An open and extensible framework for spatially explicit land use change modelling: the lulcc R package. Geosci Model Dev 8:3215–3229

National Research Council (2014) Advancing land change modeling: opportunities and research requirements. National Academies Press, Washington, D.C

Natural Resources Canada (2019) Mountain pine beetle. https://www.nrcan.gc.ca/our-natural-resources/forests-forestry/wildland-fires-insects-disturban/top-forest-insects-diseases-cana/mountain-pine-beetle/13381. Accessed 2 June 2020

Paudel S, Yuan F (2012) Assessing landscape changes and dynamics using patch analysis and GIS modeling. Int J Appl Earth Obs Geoinf 16:66–76

Pérez L, Dragićević S, White R (2013) Model testing and assessment: perspectives from a swarm intelligence, agent-based model of forest insect infestations. Comput Environ Urban Syst 39:121–135

Perez L, Molowny-Horas R, Harati S (2016) Modelling forest insect outbreaks: efforts towards an inverse approach to model calibration. In: Sauvage S, Sánchez-Pérez JM, Rizzoli AE (eds) Proceddings of the 8th International Congress on Environmental Modelling and Software (iEMSs). Toulouse, France, p 688

Pijanowski BC, Pithadia S, Shellito BA, Alexandridis K (2005) Calibrating a neural network-based urban change model for two metropolitan areas of the Upper Midwest of the United States. Int J Geogr Inf Sci 19:197–215

Pontius RG Jr (2000) Quantification error versus location error in comparison of categorical maps. Photogramm Eng Remote Sens 66:1011–1016

Pontius RG Jr (2018) Criteria to confirm models that simulate deforestation and carbon disturbance. Land 7:1–14

Pontius RG Jr (2002) Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. Photogramm Eng Remote Sensing 68:1041–1050

Pontius RG Jr, Boersma W, Castella JC et al (2008) Comparing the input, output, and validation maps for several models of land change. Ann Reg Sci 42:11–37

Pontius RG Jr, Castella J-C, de Nijs T et al (2018) Lessons and challenges in land change modeling derived from synthesis of cross-case comparisons. In: Behnisch M, Meinel G (eds) Possible urban futures: the impact of planners and developers on urban dynamics. Springer International Publishing, Cham, pp 143–164

Pontius RG Jr, Huffaker D, Denman K (2004) Useful techniques of validation for spatially explicit land-change models. Ecol Modell

Pontius RG Jr, Millones M (2011) Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int J Remote Sens 32:4407–4429

Pontius RG Jr, Parmentier B (2014) Recommendations for using the relative operating characteristic (ROC). Landsc Ecol 29:367–382

Pontius RG Jr, Peethambaram S, Castella JC (2011) Comparison of three maps at multiple resolutions: a case study of land change simulation in cho don district, Vietnam. Ann Assoc Am Geogr 101:45–62

Pontius RG Jr, Santacruz A, Tayyebi A, et al (2015) TOC: Total Operating Characteristic Curve and ROC Curve. R package version 0.0–4 https://cran.r-project.org/web/packages/TOC/index.html

Pontius RG Jr, Si K (2014) The total operating characteristic to measure diagnostic ability for multiple thresholds. Int J Geogr Inf Sci 28:570–583

Pontius RG Jr, Walker R, Yao-kumah R et al (2007) Accuracy assessment for a simulation model of Amazonian deforestation. Ann Assoc Am Geogr 97:677–695

Province of British Columbia (2015) BC MPB observed cumilative Kill - vol.12

Province of British Columbia (2020) Aerial Overview Survey Methods. https://www2.gov.bc.ca/gov/content/industry/forestry/managing-our-forest-resources/forest-health/aerial-overview-surveys/methods. Accessed 2 June 2020

R Core Team (2019) R: a language and environment for statistical computing

Rollins MG, Keane RE, Parsons RA (2004) Mapping fuels and fire regimes using remote sensing, ecosystem simulation, and gradient modeling. Ecol Appl 14:75–95

Rykiel EJ (1996) Testing ecological models: the meaning of validation. Ecol Modell 90:229–244. https://doi.org/10.1016/0304-3800(95)00152-2

Tobler WR (1970) A computer movie simulating urban growth in the detroit region. Econ Geogr 46:234

van Vliet J, Bregt AK, Brown DG et al (2016) A review of current calibration and validation practices in land-change modeling. Environ Model Softw 82:174–182

van Vliet J, Bregt AK, Hagen-Zanker A (2011) Revisiting Kappa to account for change in the accuracy assessment of land-use change models. Ecol Modell 222:1367–1375

Verburg PH, Kok K, Pontius RG Jr, Veldkamp A (2006) Modeling land-use and land-cover change. In: Lambin EF, Geist H (eds) Land-use and land-cover change. Springer, Berlin, pp 117–135

White R (2006) Pattern based map comparisons. J Geogr Syst 8:145–164