Computers and Electronics in Agriculture 105 (2014) 1–8

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Point-trained models in a grid environment: Transforming a potato late blight risk forecast for use with the National Digital Forecast Database

Kathleen Baker[a,*], Paul Roehsner[a], Thomas Lake[b], Douglas Rivet[a], Susan Benston[a], Bryan Bommersbach[a], William Kirk[c]

[a] Department of Geography, Western Michigan University, Kalamazoo, MI, USA
[b] Department of Computer Science, Western Michigan University, Kalamazoo, MI, USA
[c] Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI, USA

ARTICLE INFO

Article history:
Received 27 August 2013
Received in revised form 21 March 2014
Accepted 9 April 2014

Keywords:
Geographic information systems
Artificial neural network
NoSQL
Decision support system
Model development

ABSTRACT

As publicly available weather forecasting datasets advance in accuracy and spatial and temporal resolution, it is relatively simple to apply these established models to new datasets but the results may deviate from what users of decision support systems have come to expect. Potato late blight risk models were some of the earliest weather-based models. This analysis compares two types of potato late blight risk models that were originally trained on location specific (point) data in Michigan. A unique system using NoSQL was developed to train, validate and implement potato late blight risk modeling using a grid data format. Each model was tested two ways; it was first deployed directly with gridded weather forecasting data as a replacement for point data, and then retrained on the gridded data. Despite consistently lower overall accuracy, the grid trained artificial neural network model was deemed of better quality for use by stakeholders because of its accuracy on days with potato late blight risk. However, the success of the model was dependent upon its retraining using the newly available data source. In the direct implementation scenario without retraining, a simpler modified-Wallin model achieved better results than the neural network model.

1. Introduction

Weather-based prediction models have been used to estimate environmental conditions that are favorable for risk of agricultural crop disease epidemics and to make management recommendations appropriate to that risk for more than 60 years (Beaumont, 1947; Cook, 1949; Wallin and Schuster, 1960). Potato late blight risk models were some of the earliest weather-based models. As publicly available weather forecasting datasets advance in accuracy and spatial and temporal resolution, it is relatively simple to apply these established models to new datasets, but the results may deviate from what users of decision support systems have come to expect (Baker et al., 2014). Weather based prediction models are often adjusted or recalibrated to improve accuracy over time. Model retraining has long been shown to be important to model accuracy improvement in the development of point based models (Johnson et al., 1996; Raposo et al., 1993) and in gridded models (McBride and Ebert, 2000). Models developed as regional warning systems may need recalibration for more local applications or vice versa (Taylor et al., 2003). However, some recent papers using new gridded datasets do not describe model retraining when new datasets are employed (Pavan et al., 2011).

Access, retrieval and processing of data, as well as storage of derived outputs, have all become issues in agroecosystem forecasting. Recently there has been a push in the big data computing environment against standard relational tables applied to all data storage problems in an anti-one size fits all movement (Stonebraker and Cetintemel, 2005). This trend continues as systems engineering and architectural challenges are further understood as implemented within federal agencies, and effective analytic and data collection processes, system organization, and data dissemination practices become common at such scales (Begoli and Horey, 2012). With model advancements and update techniques, parallel advancements in computational capacity and data formats have expanded the possibilities of spatial decision support systems from traditional GIS and relational database

* Corresponding author. Address: Department of Geography, Western Michigan University, 1903 W. Michigan Ave. MS 5424, Kalamazoo, MI 49008-5424, USA. Tel.: +1 269 387 3345; fax: +1 269 387 3442.

E-mail addresses: kathleen.baker@wmich.edu (K. Baker), paul.j.roehsner@wmich.edu (P. Roehsner), thomas.l.lake@wmich.edu (T. Lake), dougr188@gmail.com (D. Rivet), susan.benston@wmich.edu (S. Benston), bryan.l.bommersbach@wmich.edu (B. Bommersbach), kirkw@msu.edu (W. Kirk).

http://dx.doi.org/10.1016/j.compag.2014.04.002
0168-1699/© 2014 Elsevier B.V. All rights reserved.

management systems to encompass scientific data formats and recent advances in NoSQL formats for processing of large datasets. NoSQL solutions, particularly, were developed in reaction to and as a solution for big data storage necessary to today's Internet applications. A running theme in the NoSQL movement is that the nature of the data to be stored and the pattern of data should be taken into account when choosing the technology to store and serve that data. Storing spatial data and indexed querying of spatial data has slowly been added to the feature set of a few of the NoSQL databases. Column-family store NoSQL databases have been used to house tiles for remote sensing image databases (Xiao and Liu, 2011) and graph network NoSQL databases have proven to be faster than PostgreSQL when querying point-to-point queries along the transportation network (Baas, 2012). Document store NoSQL formats hold great promise for standard GIS operations because they offer significant flexibility in the document paradigm as an entity can have any set of keys pointing to values of different data types including lists, records, audio–visual media, or geometries. This allows both raster and vector data to be stored in individual documents. Attributes and related metadata can be stored in fields within a document and the actual data can be stored alongside it. Since the clear association of metadata with data is a typical problem in forecasting systems, we selected a document store NoSQL database.

We compared the transition to a grid data environment for two types of potato late blight risk models that were originally trained on location specific (point) data in Michigan. The big data issues associated with manipulating gridded weather datasets provided an opportunity to use NoSQL system design to train, validate and implement potato late blight risk modeling using a grid data format. While the system design was novel, our primary purpose here was to examine the transition of standard types of plant pathology models to the grid environment. Each of the two late blight models was tested two ways; it was first deployed directly with gridded weather forecasting data as a replacement for point data, and then retrained on the gridded data. An estimation model, the modified-Wallin used operationally by Michigan State University since the late 1990s (Baker et al., 2000), and the latest version of a Neuro Weather Net model (NWN), an artificial neural network based forecasting model, for forecasting late blight risk (Baker et al., 2014) were the models tested for their efficacy and accuracy on the National Digital Forecast Database (NDFD) for Michigan. Both original point models were trained in Michigan on data from 2003 through 2008. These models were then retrained to adapt them to peculiarities of the grid data format, and the adjusted models were also tested for their efficacy and accuracy. Michigan has a highly variable late summer climate and therefore is a suitable place to test such models. Results of the models in the grid environment are compared across space and time for the 2009 through 2012 growing seasons.

2. Methods

2.1. Data and models

Gridded forecasts are available from the National Digital Forecast Database (NDFD) for the US at 5 km spatial resolutions and 1–3 h temporal resolution for 72 h in Gridded Binary (GRIB) format (National Weather Service, 2013). In this pilot study, we examine only the first forecast day (24 h forecast) available from NDFD for quality. Crop disease risk prediction model development requires use of this data in dimensionally large space (x, y) and time (z) axes for multiple growing seasons. The validation for NDFD is similarly scaled data available as Real-Time Mesoscale Analysis (RTMA) Products (National Oceanic and Atmospheric Administration, 2012).

To test the direct application of point trained models in a grid environment, two models, the modified-Wallin used operationally by Michigan State University since the late 1990s (Baker et al., 2000) and the latest version of an artificial neural network (ANN) based late blight risk forecasting model (Baker et al., 2014), were ported directly to NDFD data without additional training. Both models use standard meteorological variables to estimate (Wallin) or forecast (NWN) potato late blight disease risk in Michigan throughout the growing season (May 1st through September 30). The modified-Wallin model uses only hourly temperature and relative humidity as inputs, while the ANN forecast model incorporates nine variables derived from the extended forecast model output statistics from the US National Weather Service published each day at 00 UTC (Maloney et al., 2010). These daily variables include minimum temperature, cloud cover (AM and PM), quantity of precipitation estimates (AM and PM), hours estimated to be above the leaf wetness and temperature thresholds necessary for potato late blight development (calculated both with and without precipitation) and an hourly modified-Wallin based estimate of potato late blight risk overall conditions for the day (calculated both with and without precipitation). Variables that were available every three hours in NDFD or RTMA weather products were estimated at other hours through linear interpolation. Cloud cover variables were changed from a 3 category format, as used in the point trained model, to continuous scale sky coverage data, as available in the NDFD data.

After the initial model runs based on the direct porting of point based models to the grid environment, retraining strategies were implemented to adjust both the modified-Wallin and NWN model outputs to data characteristics of NDFD inputs. Because a data archive of only 4 years exists for NDFD data, each year (2009 through 2012) was examined using a standard correction based on the results of the other 3 years as a way to increase sample size and incorporate seasonal variability. This approach simulates a situation in which all available prior data is used to train a model for the coming year. In the case of the modified-Wallin model, consistent biases in the NDFD relative humidity variable yielded lower risk than expected. Linear regression was performed on each grid cell using hourly forecast relative humidity as an independent variable to predict actual relative humidity as recorded in the RTMA dataset. If the regression was statistically significant for a given grid cell, it was used in the adjusted model to estimate hours above the relative humidity threshold for all forecast days at that location, yielding the Wallin + RH Regression model.

To retrain the NWN artificial neural network forecast model, developing the NWN Grid Trained model, a sample of 10% of the Michigan grid cells were selected at random. This data was then further randomly split into two datasets, the first for training (70%) and the second for validation (30%). As with the point trained NWN model, the weights and biases of each ANN are initialized randomly to small values sampled from a normal distribution with mean 0 and variance 0.1. The parameters were then fit using stochastic gradient descent under the log-loss function (Buja et al., 2005). Hyperparameter values were optimized using random search (Bergstra and Bengio, 2012). A learning rate of 0.01, momentum of 0.1, and 10 hidden units were used for all experiments, as they were found to give near optimal performance across all datasets. Small variations in any of these parameters did not yield significantly reduced performance. The training procedure was repeated five times for each dataset starting from different random initial weight configurations. To prevent overfitting the validation data was used for early stopping with performance measured using an F_2-score (Eq. (1)), a weighted harmonic mean of precision and recall which balanced the importance of recall, or accuracy, on late blight risk days, with overall model accuracy on all days. Precision is defined as the percentage of correctly