Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Preventive Veterinary Medicine 164 (2019) 15–22

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

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

Forecasting herd-level porcine epidemic diarrhea (PED) frequency trends in Ontario (Canada)

Toluwalope Ajayi[a], Rozita Dara[b], Zvonimir Poljak[a,*]

[a] Department of Population Medicine, University of Guelph, Guelph, ON, Canada
[b] School of Computer Science, University of Guelph, Guelph, ON, Canada

ARTICLE INFO

Keywords:
Porcine epidemic diarrhea
Disease surveillance
Disease forecasting
Random forest
Classification and regression trees
Artificial neural networks

ABSTRACT

Porcine Epidemic Diarrhea Virus (PEDV) emerged in North America in 2013. The first case of PEDV in Canada was identified on an Ontario farm in January 2014. Surveillance was instrumental in identifying the initial case and in minimizing the spread of the virus to other farms. With recent advances in predictive analytics showing promise for health and disease forecasting, the primary objective of this study was to apply machine learning predictive methods (random forest, artificial neural networks, and classification and regression trees) to provincial PEDV incidence data, and in so doing determine their accuracy for predicting future PEDV trends. Trend was defined as the cumulative number of new cases over a four-week interval, and consisted of four levels (zero, low, medium and high). Provincial PEDV incidence and prevalence estimates from an industry database, as well as temperature, humidity, and precipitation data, were combined to create the forecast dataset. With 10-fold cross validation performed on the entire dataset, the overall accuracy was 0.68 (95% CI: 0.60 – 0.75), 0.57 (95% CI: 0.49 – 0.64), and 0.55 (0.47 – 0.63) for the random forest, artificial neural network, and classification and regression tree models, respectively. Based on the cross-validation approach to evaluating predictive accuracy, the random forest model provided the best prediction.

1. Introduction

Porcine Epidemic Diarrhea (PED) is a highly infectious swine disease which can cause high morbidity and mortality in swine populations (Carvajal et al., 2015). Its causative agent, Porcine Epidemic Diarrhea Virus (PEDV), recently emerged in North America, with the first reported case occurring in May 2013 in the United States (Hill et al., 2014). In Canada, the first case was identified in January 2014 on an Ontario swine farm, with additional cases reported in several provinces, from Manitoba on the west to Prince Edward Island on the east (Kochhar, 2014). The combined efforts of federal and provincial authorities, as well as industry organizations, were instrumental in controlling the spread of the disease in Ontario (Kochhar, 2014). Furthermore, the successful collaboration has led to a low PED prevalence situation in Ontario, with provincial disease elimination now viewed as a real possibility.

The Ontario Swine Health Advisory Board (OSHAB) is an industry organization which administers a PED surveillance program and database under its Area Regional Control and Elimination (ARC&E) project. The PED database allowed for tracing the PED status of individual herds in discrete locations, from time of initial infection until time of declaration of freedom from PEDV infection based on established criteria. As such, it was successfully utilized to obtain PED incidence and prevalence estimates for the province (Ajayi et al., 2018). With limited surveillance and disease control resources, accurate predictions for PED trends would be invaluable in allocating such resources, and fortunately, the PED database provides data which can be used for predictive purposes. Recent advances in forecasting and predictive methods have shown great promise for both human and animal medicine (Tu, 1996; Leung and Tran, 2000; Er et al., 2010; Mancia et al., 2012), and it seemed logical to explore these methods to determine the most appropriate one for PEDV prediction. Predictive accuracy of the methods could then be considered as one criterion for selecting the most suitable method for ongoing surveillance, which would contain short-term prediction of disease trends and possibly translate into prediction of resources needed to control the outbreak.

Therefore, the primary objective of this study was to determine the most accurate machine-learning approach for forecasting future PEDV trends in Ontario swine herds. Specifically in this study, random forest, classification and regression trees, and artificial neural networks were utilized. The secondary objective was to determine variables highly ranked as determinants for PEDV trends. The methods considered in

* Corresponding author at: Department of Population Medicine, Rm 207C, University of Guelph, Guelph, ON, N1G 2W1, Canada.
E-mail address: zpoljak@uoguelph.ca (Z. Poljak).

https://doi.org/10.1016/j.prevetmed.2019.01.005
Received 20 February 2018; Received in revised form 11 January 2019; Accepted 11 January 2019
0167-5877/

this study were selected for various reasons. Random forest, based on different metrics, showed slightly better performance in comparison to other methods when applied to similar datasets (Kane et al., 2014; Petukhova et al., 2018). Among the data-driven prediction methods, classification and regression trees is the most transparent method with results that are easy to interpret (Shmueli et al., 2017), a feature which may be preferred by end users. In addition, classification and regression trees form the basis for random forests. By contrast, results of analysis conducted via artificial neural networks cannot be directly interpreted through coefficients or through other means, however they have been successfully applied to the analysis of complex relationships between predictors and an outcome, and have often resulted in high prediction accuracy (Shmueli et al., 2017).

2. Methods

2.1. Data and data processing

Data about weekly incidence and prevalence measures were obtained from an industry database, which tracks the PEDV infection status of individual premises participating in an industry-driven voluntary disease control program. The program is known as the Ontario Area Regional Control and Elimination (ARC&E) project. Detailed explanation of variables representing herd-level PEDV infection status is provided elsewhere (Ajayi et al., 2018). Briefly, (i) confirmed positive premises had to have at least one RT-PCR positive test for the PED virus, (ii) presumed positive premises did not have diagnostic testing conducted but housed animals that were moved from known positive sites (e.g. nurseries supplied from known positive sow sites), (iii) presumed negative premises were previously confirmed or presumed positive sites that underwent diagnostic testing based on industry standards (at least 10% design prevalence with 95% herd sensitivity), which resulted in all negative tests, and (iv) confirmed negative premises were premises with no clinical signs or positive tests for PEDV for at least 6 months after the presumed negative status had been achieved. The source population for the outcome measurement consisted of commercial swine herds which included: 14 farrow-wean herds (9.3%), 5 farrow-feeder (3.3%), 16 farrow-finish (10.6%), 17 nurseries (11.3%), 7 wean-finish facilities (4.6%), 91 finisher sites (60.3%) and 1 isolation/acclimation unit (0.7%). The median number of sows on study sites was 750 sows (interquartile range = 925).

Furthermore, in calculating herd-level PED prevalence, all swine herds available in the database in a given week were used, with their PEDV status set as “undetermined”. These were herds that existed in the disease control database but were not tested for PEDV. However, due to the emerging nature of PEDV in Ontario during the study period, as well as reporting obligations in this phase of the epidemic, this “undetermined” status can be equated to negative PEDV status. These statuses were then processed to obtain weekly measures of disease frequency, including number of new cases, herd-level prevalence (expressed as a percentage), and number of infectious sows (expressed as raw counts). The number of new cases was of primary interest and was used to form the final outcome for the analysis, whereas herd-level prevalence and number of infectious sows were a-priori considered as potentially important predictors for the number of new cases and severity of the outbreak.

Province-level temperature and humidity values representative of swine locations in Ontario were obtained by aggregating corresponding county-level weather data, first to daily and then to weekly level. This was done by: 1) identifying Ontario counties with the highest pig counts as documented by the Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA, 2014), namely, the Perth, Huron, Middlesex, and Wellington counties; 2) locating a weather station in each of those counties using the Weather Underground website (www.wunderground.com), and 3) for each weather station, obtaining the following daily values for November 24, 2013 – May 6, 2017: (i) high, average, and low temperature; (ii) high, average, and low humidity; (iii) precipitation. If daily weather-related values were not reported for a weather station, daily values from another weather station in the same county, either from the Weather Underground website or Environment Canada's historical weather data, were substituted. Temperature, humidity and precipitation were selected for inclusion for two principal reasons. First, these measurements were readily available from the majority of weather stations. Second, temperature and humidity are reported to influence survival of different viruses (Lowen et al., 2007; Casanova et al., 2010) including PEDV (Thomas et al., 2015).

The daily weather data for each county were imported into R (R Core Team, 2017) and a singular dataset of temperature and humidity values for all 4 counties created. The representative daily high, average, and low temperature and humidity values for Ontario were obtained by averaging the corresponding values across the counties (i.e. Perth values + Huron values + Middlesex values + Wellington values divided by 4). A similar averaging was done for county precipitation values.

Once daily county values were aggregated to the provincial level using the methods noted above, there were seven variables in all: (i) ontario_hightemp – highest temperature reading in Ontario for the day, (ii) ontario_avgtemp – average temperature reading in Ontario for the day (iii) ontario_lowtemp – lowest temperature reading in Ontario for the day; (iv) ontario_highhumid – highest humidity reading in Ontario for the day; (v) ontario_avghumid – average humidity reading in Ontario for the day; (vi) ontario_lowhumid – lowest humidity reading in Ontario for the day; (vii) ontario_precip – precipitation reading in Ontario for the day.

For aggregation to the weekly level, a week was defined as beginning on Sunday and ending on Saturday. As such, for each week beginning November 24, 2013, a corresponding weekly value for each variable was generated by simply averaging the daily values (adding all the daily values in the week and dividing by 7), with the only exception being precipitation where a weekly sum (rather than an average) of precipitation values was generated.

The weekly temperature and humidity values were then combined with weekly PEDV data (incident cases, infectious sows, and prevalence) to produce the forecast dataset with 10 variables. As PEDV emerged in Canada in 2014, the starting week in the forecast dataset was set to January 5, 2014, with the ending week set to April 30, 2017. Each variable was then lagged five times (i.e. the corresponding values for prior weeks – up to five weeks in the past – were aligned with current values), with each weekly lag resulting in an additional variable. With the lags completed, the final incident cases dataset had a total of 174 observations and 60 variables.

A 4-week moving sum of incident cases (i.e. total number of new PED case herds at the province-level over a four-week period, with week beginning January 5, 2014) was calculated with the **zoo roll-apply** function in R (Zeileis and Grothendieck, 2005). A moving 4-week sum started with the current week of interest and included 3 additional prospective weeks in the future. Then, a 4-week moving sum equal to ‘0′ incident cases (i.e. zero new positive PED herds in the entire province of Ontario) was classified as trend “zero”, a moving sum equal to ‘1′ or ‘2′ incident cases was classified as trend “low”, a moving sum between ‘3′ and ‘6′ incident cases was classified as trend “medium”, while a moving sum greater than ‘6′ incident cases was classified as trend “high”.

Such classification was made with a rationale that the capacity needed to manage a production-limiting disease in the entire source population - swine herds in the province of Ontario - relies heavily on industry resources (i.e. different industry organizations and a limited number of veterinary practitioners with other daily responsibilities). The weekly count of zero is a preferred condition and qualitatively different from other outcomes; trend “low” within a 4-week window would not be considered as unusual, particularly during periods when the disease peaks seasonally due to environmental conditions (e.g. winter); trend “medium” would be considered as manageable, but

would also be reason for further investigation. Finally, greater than 6 new cases in a 4-week period (i.e. trend "high") would be considered as an alert and a potentially re-emerging scenario. It follows then that the authors were particularly interested in accurate classification of trends "zero" and "high". The weekly count of incident cases was then removed from the dataset, however, lags 1–5 of the weekly count of incident cases was retained. The final trend forecast dataset had a total of 171 observations and 60 variables. For this dataset, the response variable was set to the 4-week trend (with possible values "high", "medium", "low", and "zero"), and the explanatory variables set as the remaining 59 variables.

Descriptive statistics were generated and visually assessed by separating the time-series of incident cases and explanatory variables into long-term trend, seasonality and error components via the *stl* function (R Core Team, 2017).

2.2. Modeling approach

In this study, we aimed to predict the general trend in the number of new cases over a future 4-week window, and to subsequently evaluate the accuracy of such predictions with three different classification approaches.

2.2.1. Classification methods

The random forest, artificial neural nets, and classification tree algorithms were selected for forecasting PEDV trends. The following sections provide a brief overview of the methods.

The random forest algorithm (Breiman, 2001) is a non-parametric predictive modeling method which works by: 1) constructing multiple classification or regression trees, and 2) aggregating results from these trees to generate a prediction ('y' or response variable) for a specified set of input values ('x' or explanatory variables). Regression trees are constructed for a *continuous* response variable with the final prediction 'y' determined by averaging results across all trees. For a *categorical* response variable, classification trees are constructed and the final prediction 'y' is determined by a majority class vote across all trees. Furthermore, rather than determine the best node split by looking at *all explanatory variables* at a given node (as is the case with standard classification and regression tree algorithms), the random forest algorithm *randomly selects a subset of explanatory variables* at each node - which reduces the correlation between subtrees (Ho, 2002) - and then determines the best (or homogenous) binary split at the node.

The random forest implementation in R - *randomForest* (Liaw and Wiener, 2002) - was used in this study. It provides, amongst other features, tuning functions for ascertaining the number of explanatory variables which should be randomly sampled at each node, as well as the optimal number of variables for predicting 'y'. In addition, the random forest implementation provides a variable importance measure, which ranks each explanatory variable per the mean decrease in prediction accuracy when the variable is randomly permuted and other explanatory variables left unchanged.

Neural nets (or artificial neural networks) are predictive algorithms developed to mimic biological activity in the human brain, specifically the learning patterns for neurons. Neural networks have an input layer, hidden layer(s), and an output layer made up of interconnected neurons and an activation function. Predictors are supplied to the input layer, which transfers these values to one or more hidden layer(s) for processing via a system of weighted connections. These hidden layer(s) in turn link to an output layer which provides the final prediction result. Tuning parameters, such as the maximum number of learning iterations, learning rate, and number of hidden layers and weights, can be set for artificial neural networks (Shmueli et al., 2010).

The current neural network implementations in R are *nnet* (Venables and Ripley, 2002) and *neuralnet* (Fritsch and Guenther, 2016). The *nnet* implementation permits one hidden layer, and has several tuning parameters, such as *size* (for specifying the number of neurons in the hidden layer), *decay* (a weight decay value to aid in the model optimization process and avoid overfitting), and *maxit* (the maximum number of permitted iterations).

Classification trees are predictive algorithms which utilize recursive partitioning, a step-by-step process which splits a node into sub-nodes by evaluating a Boolean condition at each node. Observations which meet the Boolean condition are placed in one node, while the remaining observations are placed in another node. The process is repeated until a terminal node (which can no longer be split) is reached and a class label is assigned. Sub-trees are built with each recursive split, and each split (or partition) is constructed such that the resulting nodes are homogenous in nature (Izenman, 2008).

The classification tree implementation in R - *rpart* (Therneau et al., 2017) – was used in this study. The *rpart* implementation has several tuning parameters, such as the minimum number of observations which must be present in a node for a split to be attempted, the number of cross-validations performed on observations to determine the best split, as well as complexity parameter *cp*, where any split that does not decrease the overall lack of fit by a factor of *cp* is not attempted.

2.2.2. Model development on training dataset

The forecast dataset was split into 70% training and 30% testing using functionality available in the *caret* package (Kuhn, 2008). Class distribution was taken into account for the dataset splits, which ensured a balance of "high", "medium", "low", and "zero" trend observations for the training and test set. Initially, the random forest model was fitted on the training dataset with all 59 explanatory variables - the number of variables sampled at each split was obtained from the tuning function and a variable importance plot was also constructed. Going forward, this model will be known as the "full" model. Furthermore, the random forest tuning function, which provides the optimal number of variables for predicting 'y', indicated the lowest prediction error rate at 30 variables. As such, **all subsequent models,** including the random forest model, were constructed with the top 30 variables from the full model (as obtained from the variable importance plot). For each of the three methods used herein, different parameters for each modelling approach, were then fine-tuned by conducting repeated 10-fold cross-validation.

For the subsequent random forest model, the number of trees grown was set to 500, the number of variables randomly sampled at each split was 3, and a variable importance plot was also generated. For the subsequent neural nets model, the dataset was centered, and scaled, with the number of nodes in the hidden layer, decay function and number of iterations set to maximize the classification accuracy on the testing dataset. The number of nodes in the final neural nets model (i.e. size) was set to 11, decay was set to 0.2, and the maximum number of iterations was set to 1000. The neural nets algorithm used was based on the feed-forward method. For the subsequent classification trees model, the minimum number of observations for a split attempt was set to 14, and the complexity parameter was set to 0.01.

2.2.3. Evaluation of predictive accuracy for future trends

For the classification approach with 70% training and 30% test sets, model performance was assessed via confusion matrices, overall classification accuracy, as well as model sensitivity and specificity values by trend (i.e. "high", "medium, "low", and "zero"). Due to concerns about limited observations (i.e. small dataset) and to further validate the model's performance for future observations, an additional approach was used to evaluate the predictive accuracy of the models. This was based on 10-fold cross validation, which was performed once on the **entire** dataset using functionality available in the *caret* package. Specifically, the entire dataset was randomly split into 10 equal-sized groups (or folds), with one fold allocated as the test set and the remaining nine as the training set. The test and training sets were then fed to random forest, neural nets, and classification tree models, with explanatory variables and tuning parameters set to the values used

Table 1
PEDV trend classification performance on the test dataset for Ontario PEDV and weather data from January 2014 - April 2017 (171 observations - 70% of the dataset allocated for model training and 30% allocated for model testing).

	Prediction				Kappa	95% CI	Overall Accuracy	95% CI
Reference	High	Medium	Low	Zero				
training set performance - random forest model with 30 co-variates								
High	9	5	0	0	0.50	0.38 – 0.62	0.64	0.55 – 0.72
Medium	3	17	7	3				
Low	0	7	23	9				
Zero	0	2	8	29				
training set performance - neural nets model with 30 co-variates								
High	12	1	1	0	0.88	0.80 – 0.95	0.91	0.84 – 0.95
Medium	0	28	2	0				
Low	0	1	35	3				
Zero	0	0	3	36				
training set performance - classification tree model with 30 co-variates								
High	11	0	3	0	0.73	0.63 – 0.82	0.80	0.72 – 0.87
Medium	1	22	6	1				
Low	2	4	33	0				
Zero	0	3	4	32				

Table 2
PEDV trend classification performance on the test dataset for Ontario PEDV and weather data from January 2014 - April 2017 (171 observations - 70% of the dataset allocated for model training and 30% allocated for model testing).

	Prediction				Kappa	95% CI	Overall Accuracy	95% CI
Reference	High	Medium	Low	Zero				
test set performance - random forest model with 30 co-variates								
High	3	1	1	0	0.60	0.38 – 0.62	0.71	0.57 – 0.83
Medium	1	9	2	0				
Low	1	0	11	4				
Zero	0	1	3	12				
test set performance - neural nets model with 30 co-variates								
High	4	0	1	0	0.66	0.49 – 0.82	0.76	0.61 – 0.87
Medium	1	7	2	2				
Low	1	0	12	3				
Zero	0	0	2	14				
test set performance - classification tree model with 30 co-variates								
High	2	2	1	0	0.23	*0.00 – 0.35	0.45	0.31 – 0.60
Medium	2	4	3	3				
Low	1	8	4	3				
Zero	0	5	2	9				

*truncated to 0.

previously (i.e. for the corresponding models with a 70-30 split of the data). The process of allocating test and training sets was repeated until each fold had been used exactly once as a test set. Class distribution for the outcome of interest was not considered for each fold, and as such, the randomly selected training and test sets did not have balanced "high", "medium", "low", and "zero" trend observations representative of the entire dataset. For the 10-fold cross-validation approach, model performance was assessed via a summary confusion matrix with overall classification accuracy (across the 10 folds), model sensitivity and specificity values for each trend (i.e. "high", "medium, "low", and "zero"), as well as a boxplot of model sensitivity and specificity values by trend across all 10 folds. Model estimates of accuracy based on the cross-validation approach were compared using a paired *t*-test (Kuhn and Johnson, 2016), in which each fold represented the unit of observation.

3. Results

In the full dataset, 19 observations were classified as having "high" trend, 42 as "medium" trend, 55 as trend "low", and 55 as "zero" (Tables 1 and 2). For the forecasting of PEDV trends, the confusion matrix in Table 1 provides accuracy measures on the training set for a static 70-30 split of the data. Table 2 provides accuracy measures on the test dataset for the same split, with random forest, artificial neural nets, and classification trees reporting an overall accuracy of 71%, 75%, and 45% respectively. If non-tolerable errors are considered as misclassification into non-adjacent categories (e.g. high as low, high as zero, medium as zero, and vice versa), and tolerable errors considered as misclassification into adjacent categories (e.g. high as medium, medium as low, low as zero, and vice versa), then random forest had 3 non-tolerable errors and 11 tolerable ones, neural nets had 4 non-tolerable errors and 8 tolerable ones, while classification trees had 10 non-tolerable errors and 20 tolerable ones.

For the additional models constructed with random training and test sets (using 10-fold cross validation on the entire dataset), the summary confusion matrix in Table 3 indicates overall accuracy values of 68%, 57%, and 55% for random forest, neural nets, and classification trees respectively. Paired *t*-test results confirmed accuracy estimates for random forests as being different from neural nets ($p = 0.02$) and classification trees ($p < 0.01$), whereas there was no difference in accuracy estimates between neural nets and classification trees ($p = 0.68$).

Table 3
PEDV trend classification performance for Ontario PEDV and weather data from January 2014 - April 2017 (171 observations), with randomly allocated training and test sets (10-fold cross validation).

	Prediction				Kappa	95% CI	Overall Accuracy	95% CI
Reference	High	Medium	Low	Zero				
10-fold cross validation performance - random forest model with 30 co-variates								
High	11	7	1	0	0.55	0.45 – 0.65	0.68	0.60 – 0.75
Medium	2	33	5	2				
Low	1	9	34	11				
Zero	0	1	16	38				
10-fold cross validation performance - neural nets model with 30 co-variates								
High	9	6	1	3	0.39	0.29 – 0.49	0.57	0.49 – 0.64
Medium	3	21	9	9				
Low	1	8	25	21				
Zero	0	2	11	42				
10-fold cross validation performance – classification tree model with 30 co-variates								
High	9	7	2	1	0.38	0.27 – 0.48	0.55	0.47 – 0.63
Medium	2	25	11	4				
Low	3	13	29	10				
Zero	1	7	16	31				

With non-tolerable errors and tolerable errors defined as before, random forest had 5 non-tolerable errors and 50 tolerable ones, neural nets had 16 non-tolerable errors and 58 tolerable ones, while classification trees had 18 non-tolerable errors and 59 tolerable ones.

The sensitivity and specificity for all models are presented in Table 4, while the boxplot of sensitivity and specificity values across 10 folds on the entire dataset is presented in Fig. 1. The boxplot of sensitivity and specificity values indicate higher median values for the

Table 4
PEDV trend classification diagnostics for Ontario PEDV and weather data from January 2014 - April 2017, for the 30% test set and randomly allocated training and test sets (10-fold cross validation).

	High	Medium	Low	Zero
diagnostics – random forest model with 30 co-variates				
Sensitivity	0.60	0.75	0.69	0.75
Specificity	0.95	0.95	0.82	0.88
diagnostics – neural nets model with 30 co-variates				
Sensitivity	0.80	0.58	0.75	0.88
Specificity	0.95	1.00	0.85	0.85
diagnostics – classification tree model with 30 co-variates				
Sensitivity	0.40	0.25	0.50	0.56
Specificity	0.93	0.81	0.67	0.82
10-fold cross validation diagnostics – random forest model with 30 co-variates				
Sensitivity	0.58	0.79	0.62	0.69
Specificity	0.98	0.87	0.81	0.89
10-fold cross validation diagnostics – neural nets model with 30 co-variates				
Sensitivity	0.47	0.50	0.46	0.76
Specificity	0.97	0.88	0.82	0.72
10-fold cross validation diagnostics – classification tree model with 30 co-variates				
Sensitivity	0.47	0.60	0.53	0.56
Specificity	0.96	0.79	0.75	0.87

random forest model across all 10-folds, with the only exception being trends “zero” and “low” for sensitivity and specificity respectively, where neural nets had a higher median value.

The variable importance plot for the random forest model is presented in Fig. 2, while a description of variable names for the 30 explanatory variables in all classification models is provided in S-Table 1.

For random forest classification, prevalence-related variables, whether current or lagged, were the driving force for the 4-week incidence trend (Fig. 2), however, the current week’s mean low temperature (i.e. *ont_meanlowtemp*) was also highly ranked. It appears the current week's low temperature, alongside prior and current prevalence values, determine the PEDV trend 4 weeks into the future.

Time series decomposition plots for weekly PEDV incident cases and prevalence, as well as weekly low temperature, average temperature, and average humidity are available in S-Figures 2, 3, 4, 5, and 6 respectively. Each plot shows a strong seasonality component, however, there are notable differences where the trend is concerned. For example, incident cases (S-Figure 2) started with a sharp trend decrease at the beginning of 2014 but then followed with a gradual decrease, while prevalence (S-Figure 3) had a gradual trend increase at the beginning of 2014 followed by a decrease. Both low and average temperature (S-Figures 4 and 5) show a trend increase over the study period, while average humidity (S-Figure 6) started with a trend decrease but then switched to a trend increase in mid-2016. Determination of long-term trend for weather variables was, however, based on little informative data (S Figures 4–6).

4. Discussion

Porcine Epidemic Diarrhea Virus (PEDV) continues to be a costly disease for swine producers, and is still cause for concern in the United States and Canada. The weekly Swine Enteric Coronavirus Disease (SECD) Situation Report from the United States Department of Agriculture (USDA) – dated August 17, 2017 - lists 2754 premises as confirmed positive for PEDV, with a cumulative total of 3379 PEDV-positive premises since reporting began in June 2014 (USDA, 2017). As of August 22, 2017, PEDV-PCR positive cases were reported in virtually all of the lower 48 states (University of California Davis, 2017).

The PEDV incursion into Canada was less severe, and although PED is not listed as a reportable or notifiable disease by the Canadian Food Inspection Agency (CFIA, 2017), it is provincially regulated and remains a reportable disease in several provinces, including Alberta (Alberta Ministry of Agriculture and Forestry, 2014), Saskatchewan (Government of Saskatchewan, 2017), and Manitoba (Manitoba Agriculture, 2017). PEDV is no longer provincially reportable in Ontario (OMAFRA, 2017), and while Quebec is free of the virus, PEDV remains a reportable disease in the province (Quebec Ministry of Agriculture, Fisheries and Food, 2017).

While it appears PEDV is under control in most regions in Canada, the recent outbreaks in Manitoba (spanning May – August 2017) are a reminder that PEDV is still present, and underscores the importance of monitoring, surveillance, and intervention strategies in controlling and

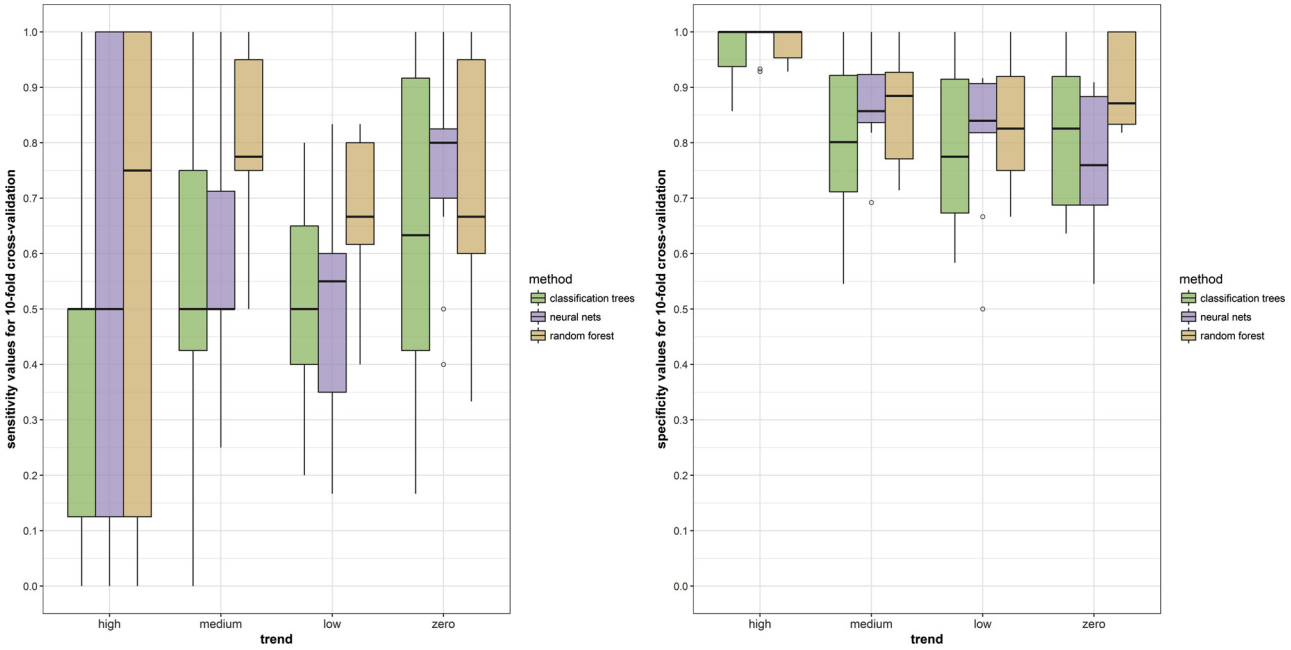

Fig. 1. Boxplots of sensitivity and specificity of PEDV trend classification across models, with 10-fold cross validation applied to the Ontario PEDV and weather dataset for January 2014 - April 2017.

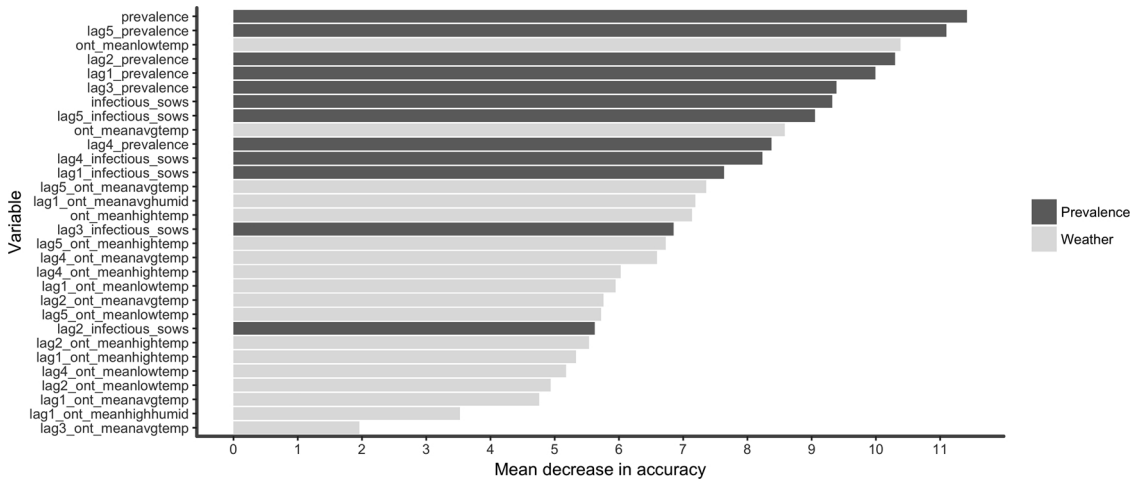

Fig. 2. Variable importance plot for the random forest classification model with 30 co-variates (PEDV long-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017. The x-axis represents the decrease in predictive accuracy once this variable has been omitted from the random forest model, with longer bars representing a larger loss in accuracy, therefore indicating the variable is of higher importance in predicting trends in the number of new cases. Variables are further colored based on whether they are related to environmental factors or level of PEDV infection.

possibly eliminating the disease. To this end, forecasting methods can be added to the surveillance toolkit as an early warning system for ongoing outbreaks.

For the above-noted PEDV forecast models with a static 70-30 split of the dataset, the artificial neural nets model was the best-performing model with the highest overall accuracy of 0.76 on the test dataset, a result which confirms one of its key strengths, which is its ability to capture complex non-linear relationships between explanatory and response variables (Shmueli et al., 2017). However, judging by an overall accuracy of 0.71 on the test dataset and the misclassification error, the random forest model is a close second. In addition, the random forest model generated just one additional misclassification error on the test dataset when compared to corresponding errors generated by neural nets, and the error was a tolerable one. In this sense, random forest performance is comparable to neural nets. Classification trees, on the other hand, was the worst-performing model, a result which confirms a key weakness of the method, which is its inability to accurately capture complex relationships between explanatory and response variables.

For the random training and test sets based on 10-fold cross validation, random forest is the best-performing model, a finding which confirms its robustness against majority-class overfitting due to an imbalanced dataset (Breiman, 2001), which in this case would be a model's tendency to classify most, if not all, observations as trend "zero" (the majority class). Given the nature of the data, there are few "high" trend observations and many "zero" trend observations. A more pronounced decrease in accuracy during cross-validation than during validation based on a split into training and validation datasets could be a consequence of the sampling approach. It is possible that randomization during cross-validation resulted in imbalanced distribution of observations across outcome categories for training or validation datasets. Alternatively, the cross-validation approach may have been a more robust approach to evaluating accuracy on new data, rather than evaluating accuracy on a single draw of randomly selected data.

In addition, while the evaluation of classification trees based on the cross-validation approach reported higher accuracy than the evaluation on a single static test dataset, it was still the worst performing model - further confirmation that this predictive method is not well-suited to the dataset. The cross-validation approach also allowed for accuracy comparisons among the three different methods via the paired *t*-test procedure. The results of this test further supported random forests as having different (i.e. higher) accuracy estimates than neural nets and classification and regression trees. Based on different metrics, results from random forests have been shown to be more accurate than several other methods on surveillance datasets of roughly comparable size (Kane et al., 2014; Petukhova et al., 2018). Nonetheless, this should not be over-interpreted since the majority of the datasets used have been moderate in size. However, what is clear from this study is that random forest, as an ensemble of classification trees, had considerable higher accuracy on the test dataset than the method that was based on a single classification tree.

In terms of model sensitivity and specificity, especially as they pertain to trends "high" and "zero", neural nets outperform all other models with a static 70-30 split of the data. However, with random training and test sets (10-fold cross-validation), it appears random forest outperforms all other models, the only exception being trends "zero" and "low" for sensitivity and specificity respectively, where neural nets had a higher median value. Since neural networks require sufficient records to "learn" the patterns for a minority class (in this case trend "high"), it is perhaps not surprising that the neural nets model focused on majority class "zero" and "low", and as such provided better sensitivity and specificity values.

Overall, the PEDV trend classification models have much higher specificity values than they do sensitivity, indicating that they're much better at identifying when a specific trend is *not* present, as opposed to when the trend is present. However, in this instance and for the purposes of surveillance for an emerging disease, the cost of a false negative (i.e. false "non-zero" trend being high), outweighs the cost of a false positive, and as such, sensitivity values are paramount.

Although the random forest classification model outperforms other models, its sensitivity value across all trends is mediocre, indicating that it may not be the best option for avoiding false negatives. It is also likely that the nature of the data puts the random forest method at a disadvantage. While random forest has shown promise for time series data related to endemic animal diseases (Kane et al., 2014), studies highlighting its use for time series data related to emerging animal diseases, especially non-zoonotic ones, are limited (Xie et al., 2016). Unlike endemic diseases, for which there are frequent infection peaks and lows, emerging diseases typically have an initial sustained peak (usually at the start of the epidemic), after which the epidemic is typically brought under control, preventive measures initiated, and case counts become low or zero. In summary, further work is needed to develop suitable models for emerging infectious disease data, as the number of observations in the dataset are likely inadequate for the training needs of machine learning models (e.g., random forest, neural nets, and classification trees).

With respect to variable importance as assessed by random forests,

among the top ten important variables, eight of them were related to existing frequency of cases and two were related to environmental conditions. This in itself is not surprising since the biggest risk factor for the number of new cases of a communicable disease is the prevalence of infection (Krämer et al., 2010). Nonetheless, prevalence and other measures of disease frequency for newly emerging diseases are often changing rapidly, as was the case for this disease (Ajayi et al., 2018), and are frequently not known with reasonable certainty. Therefore, building surveillance systems which are capable of providing up-to-date prevalence estimates for PED and similar production diseases in the source population would be valuable for making predictions about short-term disease frequency trends. This would require notification at the start of infection, as well as accurate data about declaration of freedom from infection for individual establishments.

For PEDV prediction (i.e. PEDV trends), the highly-ranked prevalence, lagged prevalence and temperature variables (from the random forest model) could be a consequence of the nature of swine production in the source population, which to a large degree is organized through segregated phases of production at various locations. For example, an outbreak in a large sow herd will eventually lead to spread of infection from a sow herd to one or more nursery sites, and eventually finisher herds. This requires varying lengths of time, depending on the organization of pig flow in affected farms. Alternatively, for swine herds not connected through pig flow to the existing cases, the length of time for PEDV transmission between existing cases and naïve herds could vary. Yet another possible contributor to the lags are reporting delays due to a variety of reasons (e.g. low clinical impact in growing pigs).

It is worth mentioning that the random forest importance measures do not necessarily indicate that the underlying associations are positive (i.e. that an increase in values of a predictor variable leads to an increase in the values of one or more outcome classes). In addition, while the above-noted situations are possible reasons lagged measures of infection and prevalence are identified as important, the reader should be reminded that in the study population, the status of nursery and finisher herds were tracked together with the status of sow herds. In fact, the majority of sites in the study and in the source population of PED positive herds consisted of sites that were housing nursery and finisher pigs (a total of 76.2% of herds). Therefore, the results may not be extrapolated directly to target populations consisting of sow herds only. However, this source population mirrors the target population of Ontario commercial herds with respect to major demographic characteristics, and both are reflective of the type of segregated production which has been dominant in commercial swine production.

Among the environmental variables, mean low temperature and mean average temperature were identified among the top ten important variables. It is paramount to note that since PEDV is more stable at low rather than high temperatures (Thomas et al., 2015), PEDV biosecurity measures are more difficult to implement in cold weather. Furthermore, it has been shown that transport vehicles play a critical role in PEDV transmission (Lowe et al., 2014), and as such, transport biosecurity guidelines are readily available (National Pork Board, 2014). However, some guidelines are harder to implement in cold weather since certain disinfectants have reduced efficacy at low temperatures (Bowman et al., 2015; EQSP, 2014), and washed transport vehicles are less likely to dry completely as freezing is more likely. The impact of low temperatures on PEDV transport biosecurity has led to the publication of cold weather disinfection guidelines (OSHAB, 2014), application of disinfectants which maintain their efficacy at low temperatures (Ferry and Benjamin, 2015), and temporary funding for costs associated with enhanced biosecurity for transport vehicles (OSCIA, 2014).

The most noteworthy limitation of the current study is the small dataset, which made evaluation of predictive accuracy challenging. In addition, classifying the number of new PED cases over a 4-week period into four classes was probably too detailed. Nonetheless, such a decision was made early in the modeling phase with the rationale that it was worthwhile exploring how the models would perform if such fragmented classification was required. In retrospect, classification into two groups would probably be more robust and permit the use of additional measures of classification accuracy. With such approaches, the use of model ensembles could have been explored in a straightforward manner. Furthermore, our decision to reduce the number of variables through the use of the random forest variable importance plot may have impacted the results, as it pre-selected variables that were potentially important for this algorithm but perhaps not as important for the other two. Nonetheless, comparison of the accuracy measures between the training and test datasets did not suggest overfitting, at least not for this specific method.

In conclusion, this study investigated the use of predictive analytics for the prediction of PEDV trends. The results show that the random forest classification model with 30 explanatory variables is the best model for forecasting future PEDV trends for this target population. Furthermore, variable importance measures from the random forest models confirm prevalence and temperature as important contributors to future PEDV trends.

Acknowledgements

The authors are thankful to Ontario Swine Health Advisory Board for sharing the data and to producers who contributed to the surveillance system. Source of funding for this work was National Science and Engineering Research Council.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.prevetmed.2019.01.005.

References

Ajayi, T., Dara, R., Misener, M., Pasma, T., Moser, L., Poljak, Z., 2018. Herd-level prevalence and incidence of porcine epidemic diarrhea virus (PEDV) and porcine deltacoronavirus (PDCoV) in swine herds in Ontario, Canada. Transbound. Emerg. Dis. 65 (1).

Alberta Ministry of Agriculture and Forestry, 2014. Swine Delta Coronavirus (SDCV) [Online]. Available at. (Accessed September 16, 2017). http://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/com14881.

Bowman, A.S., Nolting, J.M., Nelson, S.W., Bliss, N., Stull, J.W., Wang, Q., Premanandan, C., 2015. Effects of disinfection on the molecular detection of porcine epidemic diarrhea virus. Vet. Microbiol. 179, 213–218. https://doi.org/10.1016/j.vetmic.2015.05.027.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Canadian Food Inspection Agency, C.F.I.A.-, 2017. Terrestrial Animal Diseases [Online]. Available at. (Accessed September 16, 2017). http://www.inspection.gc.ca/animals/terrestrial-animals/diseases/eng/1300388388234/1300388449143.

Carvajal, A., Argüello, H., Martínez-Lobo, F.J., Costillas, S., Miranda, R., de Nova, P.J.G., Rubio, P., 2015. Porcine epidemic diarrhoea: new insights into an old disease. Porc. Heal. Manag. 1, 12. https://doi.org/10.1186/s40813-015-0007-9.

Casanova, L.M., Jeon, S., Rutala, W.A., Weber, D.J., Sobsey, M.D., 2010. Effects of air temperature and relative humidity on coronavirus survival on surfaces. Appl. Environ. Microbiol. 76 (9), 2712–2717.

EQSP - Quebec Swine Health, 2014. Useful Information on Disinfectants After Contamination With Novel Swine Enteric CoronaVirus Diseases (SECD). Retrieved from. Longueuil, Québec (Accessed September 16, 2017). https://www.opic.on.ca/images/pdfs/disinfectants/2014EQSPsummarydisinfectantseffectiveagainstPEDV.pdf.

Er, O., Yumusak, N., Temurtas, F., 2010. Chest diseases diagnosis using artificial neural networks. Expert Syst. with Appl. 37, 7648–7655. https://doi.org/10.1016/j.eswa.2010.04.078.

Ferry, B., Benjamin, M., 2015. Michigan State University Extension, Pork Producers Have Another Option for Disinfecting Against PEDV [Online]. Available at. (Accessed September 8, 2017). http://msue.anr.msu.edu/news/pork_producers_have_another_option_for_disinfecting_against_pedv.

Fritsch, S., Guenther, F., 2016. Neuralnet: Training of Neural Networks. R Package Version 1.33. https://CRAN.R-project.org/package=neuralnet.

Government of Saskatchewan, 2017. Report a Notifiable Livestock Disease [Online]. Available at. (Accessed September 16, 2017). https://www.saskatchewan.ca/business/agriculture-natural-resources-and-industry/agribusiness-farmers-and-ranchers/livestock/animal-health-and-welfare/notifiable-disease-list.

Hill, C., Raizman, E., Snider, T., Goyal, S., Torremorell, M., Perez, A.M., 2014. Emergence of porcine epidemic diarrhoea in North America. Focus on 9, 1–8.
Ho, T.K., 2002. A data complexity analysis of comparative advantages of decision forest constructors. Pattern Anal. Appl. 5 (2), 102–112. https://doi.org/10.1007/s100440200009.
Izenman, A.J., 2008. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, 1st edn. Springer Publishing Company, Incorporated.
Kane, M.J., Price, N., Scotch, M., Rabinowitz, P., 2014. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics 15, 276. https://doi.org/10.1186/1471-2105-15-276.
Kochhar, H.S., 2014. Porcine epidemic diarrhea in Canada: an emerging disease case study. Can. Vet. J. = La Rev.Vet. Can. 55, 1048–1049.
Krämer, A., Akmatov, A., Kretzschmar, M., 2010. Chapter 5. principles of infectious disease epidemiology. In: Krämer, A., Kretzschmar, M., Krickeberg, K. (Eds.), Modern Infectious Disease Epidemiology. Concepts, Methods, Mathematical Models, and Public Health. Springer.
Kuhn, M., 2008. Building predictive models in r using the caret package. J. Stat. Softw. 28, 1–26.
Kuhn, M., Johnson, K., 2016. Applied Predictive Modelling. Springer.
Leung, P., Tran, L.T., 2000. Predicting shrimp disease occurrence: artificial neural networks vs. logistic regression. Aquac. 187, 35–49. https://doi.org/10.1016/S0044-8486(00)00300-8.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.
Lowe, J., Gauger, P., Harmon, K., Zhang, J., Connor, J., Yeske, P., et al., 2014. Role of transportation in spread of porcine epidemic diarrhea virus infection, United States. Emerging Infectious Disease Journal 20 (5), 872. https://doi.org/10.3201/eid2005.131628.
Lowen, A.C., Mubareka, S., Steel, J., Palese, P., 2007. Influenza virus transmission is dependent on relative humidity and temperature. PLoS Pathog. 19 (10), 1470–1476.
Mancia, A., Ryan, J.C., Chapman, R.W., Wu, Q., Warr, G.W., Gull, F.M.D., Van Dolah, F.M., 2012. Health status, infection and disease in California sea lions (*Zalophus californianus*) studied using a canine microarray platform and machine-learning approaches. Dev. Comp. Immunol. 36, 629–637. https://doi.org/10.1016/j.dci.2011.10.011.
Manitoba Agriculture, 2017. Porcine Epidemic Diarrhea (PED) Virus [Online]. Available at. (Accessed September 9, 2017). https://www.gov.mb.ca/agriculture/animals/animal-health/porcine-epidemic-diarrhea.html.
National Pork Board, 2014. PEDV Brings Its Worst. Pork Checkoff Brings Its Best. Pork Checkoff, Des Moines, Iowa [Online]. Available at. (Accessed September 9, 2017). http://www.pork.org/wp-content/uploads/2013/11/pedvbookjan16final.pdf.
OMAFRA - Ontario Ministry of Agriculture Food and Rural Affairs, 2017. Livestock Disease Control and Prevention - Ontario-specific Immediately Notifiable Diseases [Online]. Available at. (Accessed September 16, 2017). http://www.omafra.gov.on.ca/english/livestock/vet/disease_pre.html.
OMAFRA - Ontario Ministry of Agriculture Food and Rural Affairs, 2014. Number of Pigs, by County, July 2014 [Online]. Available at. (accessed August 31, 2017). http://www.omafra.gov.on.ca/english/stats/livestock/ctypigs14.htm.
OSCIA - Ontario Soil and Crop Improvement Association, 2014. PED Program Deadline Nears and Pork Industry Tightens Biosecurity [Online]. Available at. (Accessed August 8, 2017). http://www.ontariosoilcrop.org/blog/2014/03/05/ped-program-deadline-nears-and-pork-industry-tightens-biosecurity/.
OSHAB - Ontario Swine Health Advisory Board, 2014. Cold Weather Trailer Disinfection Procedure [Online]. Available at:. (Accessed September 16, 2017). http://www.ontariopork.on.ca/Portals/0/Docs/Production/disease/OSHAB-Cold-Weather-Trailer-Disinfection-final.pdf.
Petukhova, T., Ojkic, D., McEwen, B., Deardon, R., Poljak, Z., 2018. Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada. PLoS One 13 (6), e0198313. https://doi.org/10.1371/journal.pone.0198313.
Quebec Ministry of Agriculture Fisheries and Food, 2017. Porcine Epidemic Diarrhea (DVS) and Porcine Deltacoronavirus (DCVP) [Online]. Available at. (Accessed September 16, 2017). http://www.mapaq.gouv.qc.ca/fr/Productions/santeanimale/maladies/soussurveillance/DEP/Pages/DEP.aspx.
R Core Team, 2017. R: a Language and Environment for Statistical Computing. URL. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Shmueli, G., Bruce, P.C., Inbal, Y., Patel, N.R., Lichtendahl, K.C., 2017. Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. Wiley.
Therneau, T.M., Atkinson, B., Ripley, B., 2017. rpart: Recursive Partitioning and Regression Trees. R package version 4, pp. 1–11. https://CRAN.R-project.org/package=rpart.
Thomas, P.R., Karriker, L.A., Ramirez, A., Zhang, J., Ellingson, J.S., Crawford, K.K., Bates, J.L., Kristin, J., Holtkamp, D.J., 2015. Evaluation of time and temperature sufficient to inactivate porcine epidemic diarrhea virus in swine feces on metal surfaces. J. Sw. Health Production 23, 84–90.
Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 49, 1225–1231. https://doi.org/10.1016/S0895-4356(96)00002-9.
University of California Davis, 2017. Disease BioPortal, BioPortal | Dashboard - Trends in US PEDV and PDCoV Diagnostic Data [Online]. Available at. (Accessed September 16, 2017). https://www.aasv.org/aasvwebsite/Resources/Diseases/PorcineEpidemicDiarrhea.php#BioPortalTrends.
USDA - United States Department of Agriculture Animal and Plant Health Inspection Service, 2017. Swine Enteric Coronavirus Disease (SECD) Weekly Situation Report [Online]. Available at. (Accessed September 16, 2017). http://www.aphis.usda.gov/animal-health/secd.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics With S. Fourth. Springer, New York.
Xie, G.Y., Olson, D.H., Blaustein, A.R., 2016. Projecting the global distribution of the emerging amphibian fungal pathogen, *Batrachochytrium dendrobatidis*, based on IPCC climate futures. PLoS One 11. https://doi.org/10.1371/journal.pone.0160746.
Zeileis, A., Grothendieck, G., 2005. Zoo: S3 infrastructure for regular and irregular time series. J. Stat. Software 14, 1–27 DOI: 10.18637/.