Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) in Ontario: Incidence, Prevalence, and Forecasting of Cases and Trends

by

Toluwalope Ajayi

A Thesis presented to
The University of Guelph

In partial fulfilment of requirements for the degree of Master of Science in Population Medicine

Guelph, Ontario, Canada

© Toluwalope Ajayi, November, 2017
ABSTRACT

Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) in Ontario: Incidence, Prevalence, and Forecasting of Cases and Trends

Toluwalope Ajayi
University of Guelph, 2017

Advisors:
Dr. Zvonimir Poljak
Dr. Rozita Dara

This thesis is an investigation into measures of disease frequency for Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) in Ontario, as well as outbreak forecasts for Porcine Epidemic Diarrhea Virus (PEDV) in Ontario. Surveillance data from an industry-administered voluntary disease control program (DCP) were analysed to obtain estimates of incidence risk, incidence rate, and prevalence for Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) between 2014 and 2016. The results indicate an overall decline in disease frequency measures, from a prevalence of 4.36% to 1.35% and 0.48% to 0.16% from 2014 to 2016, for PEDV and PDCoV respectively. For PEDV outbreak forecasts, incidence as well as prevalence and weather data were trained on Classification and Regression Trees (CART), Random Forest, and Artificial Neural Network models. The Random Forest model provided the best prediction for long-term PEDV trends, with the variable importance measure pointing to prevalence and low temperature as strong determinants for future PEDV outbreaks.
ACKNOWLEDGEMENTS

I would like to thank my advisors, Dr. Poljak and Dr. Dara, for supporting my graduate career and for their guidance throughout this project. A special note of gratitude goes out to Dr. Poljak, who despite unfortunate and grievous personal circumstances, still managed to make himself available to assist in every way possible - the cloud of “Epi Confusion” was apparent on many days, and it was so easy to get lost in the risks, rates, sensitivities, specificities, and positive and negative predictive values, however, he was always there to nudge me towards the “eureka” moment – “ah ha, so that’s how it is… I get it now”.

A big thanks to Dr. Andria Jones-Bitton and Dr. David Pearl, for introducing me to the world of epidemiology and guiding me through the mind-boggling maze of analytical epidemiology, and to my undergraduate professors, Dr. Robert Gann, Dr. Lawrence Nienart, and Dr. Susan Carbone, for developing my analytical skills and providing a firm foundation for graduate studies.

I would also like to thank the PopMed swine research group, especially Dylan, Danielle, Emily, Chris, Karen, and Glen, for taking in a novice and answering my pig questions – those “oinky” animals sure are interesting.

A sincere thanks to my industry and government partners - Tim Pasma from the Ontario Ministry of Agriculture Food and Rural Affairs, Marty Misener and Lori Moser from the Ontario Swine Health Advisory Board – for providing insights on swine surveillance in Ontario.

Last but not least, I am grateful to family and close friends for their advice, listening ear, prayers, and encouragement, and to my Heavenly Father who’s always watching, I am eternally grateful.
STATEMENT OF WORK

The Ontario Swine Health Advisory Board (OSHAB) provided PEDV and PDCoV surveillance data for this project. Lori Moser from OSHAB provided metadata information for fields in the OSHAB database.

The data retrieval and import into R was done by Toluwalope Ajayi, using SQL Server Reporting Services (SSRS) and an ODBC package.

The incidence risk and rate equations were generated by Dr. Zvonimir Poljak.

Data processing and analysis of measures of disease frequency were coded in R by Toluwalope Ajayi, with input from Dr. Zvonimir Poljak.

The PEDV forecast dataset, with incidence, prevalence, and weather variables, was generated by Toluwalope Ajayi.

Model development, training, and testing for Classification and Regression Trees (CART), Random Forest, and Artificial Neural Network (ANN) models were coded in R by Toluwalope Ajayi, with Dr. Rozita Dara providing input on model training, tuning, and testing, and Dr. Zvonimir Poljak providing assistance with analysis of the results.

All visualizations were done by Toluwalope Ajayi, with input from Dr. Zvonimir Poljak and Dr. Tatiana Petukhova.
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>STATEMENT OF WORK</td>
<td>iv</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER ONE – Introduction, Literature Review, and Objectives</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Literature Review</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 PEDV and PDCoV</td>
<td>2</td>
</tr>
<tr>
<td>1.2.2 Animal Disease Surveillance</td>
<td>9</td>
</tr>
<tr>
<td>1.2.3 Predictive Modelling</td>
<td>16</td>
</tr>
<tr>
<td>1.3 Objectives</td>
<td>28</td>
</tr>
<tr>
<td>References</td>
<td>30</td>
</tr>
<tr>
<td>CHAPTER TWO – Herd-Level Prevalence and Incidence of Porcine Epidemic</td>
<td></td>
</tr>
<tr>
<td>Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) in Swine</td>
<td></td>
</tr>
<tr>
<td>Herds in Ontario, Canada</td>
<td></td>
</tr>
<tr>
<td>Bookmark not defined.</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>42</td>
</tr>
<tr>
<td>Introduction</td>
<td>44</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>46</td>
</tr>
<tr>
<td>Data sources</td>
<td>46</td>
</tr>
<tr>
<td>Herd-level PEDV and PDCoV infection status definition</td>
<td>47</td>
</tr>
<tr>
<td>Weekly measures of disease frequency</td>
<td>49</td>
</tr>
<tr>
<td>Surveillance components for PEDV</td>
<td>51</td>
</tr>
<tr>
<td>Results</td>
<td>52</td>
</tr>
<tr>
<td>Surveillance components for PEDV</td>
<td>52</td>
</tr>
</tbody>
</table>
Yearly incidence and prevalence measures ........................................ 53

Weekly incidence and prevalence measures ...................................... 53

Discussion .................................................................................... 54

Acknowledgements ........................................................................ 59

References .................................................................................... 60

CHAPTER THREE – Forecasting Herd-Level PEDV Incident Cases and Trends in Ontario ... 73

3.0 Abstract .................................................................................. 73

3.1 Introduction ............................................................................ 74

3.2 Methods .................................................................................. 75

3.2.1 Data and data processing ..................................................... 75

3.2.2 Modeling approach .............................................................. 78

3.3 Results ................................................................................... 83

3.4 Discussion ............................................................................. 87

3.5 Conclusion ............................................................................ 92

References ................................................................................... 93

Tables ......................................................................................... 97

Figures ......................................................................................... 101

Supplementary Material .................................................................. 106

CHAPTER FOUR – Research Summary and Conclusions ......................... 114

References ................................................................................... 118
LIST OF TABLES

Table 2.1: Herd-level incidence risk and rate of two novel porcine coronaviruses (PEDV and PDCoV) in Ontario swine herds between 2014 and 2016, and estimated prevalence of positive cases at the end of each year based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093) ........................................................................................................................................................................63

Table 3.1: PEDV trend classification performance for Ontario PEDV and weather data from January 2014 - April 2017 (171 observations), with 70% of the dataset allocated for model training and 30% allocated for model testing ..................................................................................................................................................97

Table 3.2: 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) for model training and testing .................................................................................................................................98

Table 3.3: PEDV trend classification diagnostics for Ontario PEDV and weather data from January 2014 - April 2017, with 70% training and 30% test sets, as well as randomly allocated training and test sets (10-fold cross validation) for model training and testing .........................................................................................99

Table 3.4: A description of the 30 explanatory variables used in all PEDV trend classification models (PEDV long-term prediction), as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017 .........................................................................................................................................100
LIST OF FIGURES

Figure 2.1: Venn diagram representing overlap among the three surveillance components for identification of new cases of the Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) between 2014 and 2016 – The Animal Health Laboratory (AHL), Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA), and the Ontario Swine Health Advisory Board (OSHAB). The size of individual circles should be proportional to the number of positive cases recorded by a specific surveillance component. The overlapping areas should represent cases that were identified by as many surveillance components as are being overlapped. The numbers listed in a specific area represent number of cases that are identified by a specific surveillance component, or combination thereof. ................................................................. 64

Figure 2.2: Cumulative weekly confirmed and presumed positive cases of Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093) ................................................. 65

Figure 2.3: Cumulative weekly confirmed and presumed positive cases for Porcine Deltacoronavirus (PDCoV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093) ................................................. 66

Figure 2.4: Number of Herds with Confirmed Positive, Presumed Positive, Presumed Negative, or Confirmed Negative status for Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093) ................................................. 67

Figure 2.5: Percentage of Herds with Confirmed Positive, Presumed Positive, Presumed Negative, or Confirmed Negative status for Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093) ................................................. 68

Figure 2.6: Number of Herds with Confirmed Positive, Presumed Positive, Presumed Negative, or Confirmed Negative status for Porcine Deltacoronavirus (PDCoV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093) ................................................. 69

Figure 2.7: Percentage of Herds with Confirmed Positive, Presumed Positive, Presumed Negative, or Confirmed Negative status for Porcine Deltacoronavirus (PDCoV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093) ................................................. 70

S.Figure 2.1: Weekly incidence risk for Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board
S.Figure 2.2: Weekly incidence risk for Porcine Deltacoronavirus (PDCoV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093). 71

Figure 3.1: 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.......................... 101

Figure 3.2: Variable importance plot for the random forest regression model with all co-variates (PEDV short-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017................................................................. 102

Figure 3.3: 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 ......................................................................................... 103

Figure 3.4: Variable importance plot for the neural nets classification model with 30 co-variates (PEDV long-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017 ......................................................................................... 104

Figure 3.5: Variable importance plot for the classification tree model with 30 co-variates (PEDV long-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017 ......................................................................................... 105

S-Figure 3.1: Scatterplot of actual vs predicted weekly PEDV counts for the random forest regression model with all co-variates (PEDV short-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017.......................... 106

S-Figure 3.2: An alternate view of variable importance (rendered by the caret VarImp function) for the random forest regression model with all co-variates (PEDV short-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017 ................. 107

S-Figure 3.3: An alternate view of variable importance (rendered by the caret VarImp function) 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 ................. 108

S-Figure 3.4: Time series decomposition plot for the weekly count of PEDV incident cases, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017 .................. 109

S-Figure 3.5: Time series decomposition plot for the weekly PEDV prevalence, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017 ........................................ 110

S-Figure 3.6: Time series decomposition plot for the weekly low temperature, as obtained from
the Ontario PEDV and weather dataset for January 2014 - April 2017................................. 111

S-Figure 3.7: Time series decomposition plot for the weekly average temperature, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017................................. 112

S-Figure 3.8: Time series decomposition plot for the weekly average humidity, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017............................................. 113
CHAPTER ONE – Introduction, Literature Review, and Objectives

1.1 Introduction

An increase in the global demand for meat and meat products has translated to rapid growth in global pig production - from 1985 to 2010, global pork production grew by more than 80%, and has since become the main meat production sector (Fournie et al., 2012). Furthermore, pork production systems have undergone major changes in order to keep up with demand. Such changes, including increased pig density on farms and reduced time-to-slaughter, raise the risk of emerging pathogens in swine populations (Fournie et al., 2012). In total, 186 new swine pathogen variants were identified between 1985 and 2010, all of which belonged to 98 emerging pathogen species, 77 of which were newly found to infect pigs. Of the 98 emerging pathogen species, 43% were found to be zoonotic (Fournie et al., 2012). The rapid increase in the number of emerging pathogens requires a vigilant approach to animal health and surveillance, which by extension feeds into a larger public health framework given their zoonotic potential.

Endemic swine pathogens such as Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) continue to circulate in North-American swine populations, causing considerable impact on animal health and resulting in substantial economic losses for swine producers. In the following sections, a brief overview of two emerging swine pathogens - Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) – is presented. The role of surveillance systems as they pertain to emerging diseases and animal health is also reviewed, in addition to a brief overview of global and North American disease surveillance systems. The utility of predictive analytics for health forecasting, both from the animal and human health perspectives, are also reviewed. Finally, objectives for the thesis as well as a synopsis of each chapter, are presented.
1.2 Literature Review

1.2.1 PEDV and PDCoV

1.2.1.1 PEDV and PDCoV etiology

Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) belong to the family *Coronaviridae*, enveloped viruses which possess a positive-sense single-stranded RNA (ribonucleic acid) genome (Jung and Saif, 2015). There are four genera in *Coronaviridae*, namely, *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus* (Jung and Saif, 2015). PEDV is a member of the genus *Alphacoronavirus*, and shares this genus with other coronaviruses infectious to swine, such as transmissible gastroenteritis coronavirus (TGEV) and porcine respiratory coronavirus (PRCV). PDCoV is a member of the genus *Deltacoronavirus* (Carvajal et al., 2015).

PEDV has a diameter range of 95 - 190 nm, which includes projections approximately 18 nm long (Pensaert and de Bouck, 1978), while PDCoV has a diameter range of 60–180 nm, excluding the projections (Ma et al., 2015). PEDV is stable at low temperatures (pH 5.0–9.0 at 4°C) with reduced infectivity at high temperatures (pH 6.5–7.5 at 37 °C) (EFSA Panel on Animal Health and Welfare, 2014). The virus can be inactivated by acidic or alkaline disinfectants, when applied for a defined period at temperatures greater than 37 °C (Jung and Saif, 2015).

1.2.1.2 PEDV and PDCoV infection and disease in individual animals

PEDV and PDCoV are highly infectious swine pathogens which can cause anorexia, diarrhea, vomiting, and dehydration in pigs (Jung and Saif, 2015). While both pathogens share similar clinical signs, PEDV infections are more severe, resulting in near 100% mortality for neonatal piglets and high morbidity for older pigs (Jung and Saif, 2015). PEDV spreads via the fecal-oral route (Jung and Saif, 2015), with clinical signs appearing 1 to 4 days post-infection.
and viral shedding typically lasting 7 – 10 days (Carvajal et al., 2015; Madson et al., 2014). In some animals, PEDV RNA has been found up to 42 days post-infection, however infectious PEDV capable of horizontal transmission was only shed 14 – 16 days post-infection (Crawford et al., 2015). According to experimental studies, PDCoV clinical signs appear 2 – 3 days post-infection (EFSA Panel on Animal Health and Welfare, 2014).

1.2.1.3 PEDV and PDCoV transmission within and between populations

The predominant route of transmission between individual animals for both viruses is the fecal-oral route. At the population level, several transmission mechanisms have been documented. PEDV has been found to spread via aerosolized PEDV particles (Alonso et al., 2014), contaminated fomites such as transport vehicles and feed (Lowe et al., 2014; Dee et al., 2014), as well as through older pigs with subclinical PEDV infection (Jung and Saif, 2015). PEDV RNA has also been detected in sow milk and semen samples, however, fecal contamination of these samples cannot be ruled out (Li et al., 2012; Sun et al., 2012; European Food Safety Authority, 2014; Carvajal et al., 2015). Given the similarities in clinical signs and oft-reported co-infection with PEDV when PDCoV is present, it is likely that the transmission mechanism reported for PEDV also applies to PDCoV (Jung et al., 2016).

1.2.1.4 Historical perspective

PEDV was first identified in the United Kingdom and Belgium in the 1970s (Wood, 1977; Pensaert and de Bouck, 1978), while PDCoV was first reported in Hong Kong in 2012 (Woo et al., 2012). Following its initial identification, PEDV outbreaks were reported in several European countries in the seventies and eighties, including Belgium, Bulgaria, France, Germany, Hungary, Netherlands, Spain, Switzerland, and the United Kingdom (Popischil et al., 2002; Saif et al., 2012; Jung et al., 2015). For reasons unknown, there was a marked decrease in European
PEDV outbreaks in the nineties and subsequent years, with isolated outbreaks reported in Italy, the Czech Republic, Hungary, Spain, and the United Kingdom (Carvajal et al., 1994; Nagy et al., 1996; Pritchard et al., 1999; Rodak et al., 2004; Rodak et al., 2005; Martelli et al., 2008). Since 2010 and onwards, additional PEDV outbreaks have been reported in Belgium, France, Germany, Italy, Netherlands, and Ukraine (Henniger and Schwarz, 2014; Hanke et al., 2015; Staedler et al., 2015; Alborali et al., 2014; Sozzi et al., 2014; Van der Wolf et al., 2015; Theuns et al., 2015; Grasland et al., 2015; Beek, 2015). In Asia, PEDV was first reported in China and Japan in 1983 (Xuan et al., 1984; Takahashi et al., 1983), and then spread to Korea, the Philippines, and Thailand in the 1990s (Kweon et al., 1993; Srinuntapunt et al., 1995). Taiwan and Vietnam reported an outbreak in 2007 and 2009 respectively, while in late 2010, a severe outbreak was reported in China, resulting in a 50% – 90% mortality rate in nursing piglets (Puranaveja et al., 2009; Duy et al., 2011; Li et al., 2012; Sun et al., 2012). Following its initial identification in Hong Kong in 2012, surveillance studies also revealed the presence of PDCoV in South Korea and Thailand, in 2014 and 2015 respectively (Jung et al., 2016; Zhang, 2016).

While PEDV and PDCoV have been present in Asia and Europe for several years, both viruses are considered emerging pathogens in the Americas. In North America, the first case of PEDV was reported in the United States in April 2013, when an Ohio swine farm tested positive for the virus (Stevenson et al., 2013). Mexico reported a PEDV outbreak in July 2013 (Trujillo-Ortega et al., 2016; Fajardo et al., 2014), while in South America, PEDV outbreaks were reported in Peru (2013 – 2014) (More-Bayona et al., 2014; Quevedo-Valle, 2014), the Dominican Republic (2013 – 2014; Gómez, 2014), Colombia (2014; Rativa, 2014), and Ecuador (2014; OIE, 2014a). PDCoV was first reported in the United States in February 2014 and in Canada in March 2014, when swine farms in Ohio and Ontario respectively, tested positive for
the virus (Wang et al., 2015; Ontario Ministry of Agriculture Food and Rural Affairs, 2014). In addition, while outbreaks in North America are ongoing, reports from government ministries as well as OIE (World Organization for Animal Health) disease notification reports for the affected South American countries seem to indicate an end to PEDV/PDCoV outbreaks in the region. The reports point to high morbidity and case-fatality rates for animals in the affected countries, specifically Ecuador at 12.29% and 77.78% (OIE, 2014a), the Dominican Republic at 12.67% and 66.77% (OIE, 2014b), and Colombia at 17.94% and 31.67% (OIE, 2014c). The morbidity and mortality rate for Peru was 2% and 12% respectively (Quevedo-Valle, 2014).

The disease situation in Mexico is now considered endemic, as indicated by the final report submitted to the OIE in 2016 (OIE, 2016). By mid-year 2013, PEDV had spread to more than 200 herds in thirteen states in the United States, (Scott et al., 2016), while by April 2014 (a year after introduction), the U.S. swine industry had lost approximately 10% of its swine population to PEDV (Jung and Saif, 2015). In April 2014, the United States Department of Agriculture (USDA) convened an investigative group to consider the possible pathways of PEDV introduction into the United States. The group concluded that of all plausible scenarios, PEDV-contaminated flexible intermediate bulk containers (i.e. “feed totes” for feed transport), provided the simplest explanation for the introduction of PEDV into the country (Scott et al., 2016).

The USDA has since grouped PEDV and PDCoV under one umbrella as Swine Enteric Coronavirus Diseases (SECD), and the outbreaks have unfortunately continued. The most recent USDA SECD report, generated weekly since reporting began on June 5, 2014, indicate a cumulative total of 3,379 confirmed PEDV-positive premises, 256 confirmed PDCoV-positive premises, and 256 confirmed co-infected premises (PEDV and PDCoV) for the week of August
Of the confirmed positive premises, 506 of 3,379 have attained PEDV-negative status, 20 of 256 have attained PDCoV-negative status, and 10 of the 256 co-infected premises have attained negative status (USDA, 2017a). Furthermore, the University of California Davis provides a Disease BioPortal dashboard with weekly case counts and a map of the PEDV and PDCoV spread across the United States. For the week of August 17, 35 PEDV-PCR positive cases and 1 PDCoV-PCR positive case were reported across the country (University of California Davis, 2017).

**1.2.1.4.1 Canadian Outbreak**

Canada reported its first PEDV case on an Ontario swine farm in January 2014 (Ojkic et al., 2015). In addition to Ontario, a few additional provinces also reported PEDV infection: Prince Edward Island reported an isolated case of PEDV in February 2014, while Manitoba and Quebec reported PEDV outbreaks in February 2014 and April 2014, respectively (Kochhar, 2014). PEDV was detected in Saskatchewan in July 2014 when a transport truck tested positive for the virus (Saskatchewan Pork, 2014), while samples from a pig handling facility in Alberta tested positive for the virus in October 2014 (Alberta Ministry of Agriculture and Forestry, 2014a). By November 2014, a total of 67 farms across Canada were infected with PEDV (Kochhar, 2014). In March 2014, PDCoV cases (also classified as Swine Deltacoronavirus - SDCV), were reported on six Ontario swine farms (Ontario Ministry of Agriculture Food and Rural Affairs, 2014), and while Alberta has no reported cases of PDCoV, the virus was detected in a few samples from pig-handling facilities in May 2014 (Alberta Ministry of Agriculture and Forestry, 2014b).

Furthermore, while Quebec has reported no additional cases of PEDV since April 2015 (Quebec Ministry of Agriculture, Fisheries and Food, 2017), Manitoba has reported PEDV cases
from February 2014 until present. Specifically, as of September 1, 2017, Manitoba had 83 confirmed PED-positive premises (36 finisher herds, 31 sow herds, 16 nursery herds), with just 9 of the 83 premises having attained presumed negative status. Furthermore, while the province had a hiatus in PEDV cases from February 2015 – April 2016 and October 2016 – April 2017, the province experienced a severe outbreak in May 2017 which is still ongoing - of the 83 confirmed positive premises noted previously, 73 (or roughly 88%) were reported between May 2017 and August 2017 alone (Manitoba Agriculture, 2017).

A 2014 outbreak study by Pasma et al. found that in Ontario, most PEDV cases reported between January and July 2014 occurred in the southern counties of the province, specifically, Bruce, Huron, Middlesex, Oxford, and Perth counties (Pasma et al., 2016). In addition, a Statistical Distance Test revealed that case farms were significantly clustered in relation to controls for the 1st and 4th nearest neighbor, an indication that case farms were more likely to neighbor cases than controls. The epidemic curve for the outbreak showed an initial peak (week ending February 15, 2014) followed by a second peak (week ending March 1, 2014), a pattern which suggests a point source and secondary mode of transmission for the PEDV outbreak in Ontario. Furthermore, the virus was likely introduced to the province via feed, specifically PEDV-contaminated spray-dried porcine plasma (SDPP). Since the contaminated feed was withdrawn from the Canadian market on February 9, 2014, the initial peak on the epidemic curve could be attributed to farms exposed to the feed, while the second peak likely reflects secondary transmission due to pig movement within the production system (Pasma et al., 2016).

1.2.1.5 PEDV and PDCoV control and elimination measures

Several vaccines have been developed to counter PEDV infection following the recent outbreaks in North America. Specifically, the following vaccines are at various stages of
production: Harrisvaccines from Iowa State University (2013), Zoetis (2014), and VIDO (Vaccine and Infectious Disease Organization)-InterVac from the University of Saskatchewan (2016) (Gerdts and Zakhartchouk, 2016). While these vaccines have shown promise in reducing PEDV clinical signs and mortality in clinical trials, their efficacy in the field have not been tested (Carvajal et al., 2015). As such, prevention is still considered the best line of defense against PEDV, with an emphasis on active surveillance and strict biosecurity measures (Carvajal et al., 2015).

Industry groups and government organizations have published detailed biosecurity guidelines for PEDV. The American Association of Swine Veterinarians (AASV) has information on the use of disinfectants, as well as biosecurity guidelines for swine farms, pig-handling facilities, transport vehicles, manure processors, and swine exhibitors (American Association of Swine Veterinarians, 2017). As some disinfectants lose their efficacy in cold weather, the Ontario Pork Industry Council (OPIC) has also provided a cold weather disinfection procedure for transport trucks (Ontario Pork Industry Council, 2017). Furthermore, the Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA) has published PEDV cleaning procedures for provincial and federal processors (Ontario Ministry of Agriculture Food and Rural Affairs, 2017). In addition, a recent study evaluating PEDV biosecurity procedures has found that such procedures are effective for lowering the risk of PEDV transmission between pigs (Kim et al., 2017).

For herds already infected with PEDV, elimination of the virus involves “closing” the herd (i.e. no new animals are introduced), actively infecting the entire herd with PEDV to develop herd immunity, and sanitation of the premises when clinical signs subside (Geiger and Connor, 2013). Fecal or intestinal material from infected piglets are preferred for maximum herd
exposure to the virus (Schwartz et al., 2013), while a strict all-in, all-out rotation of animals should be observed once clinical signs subside, with thorough cleaning, disinfecting, and drying between groups. To confirm that PEDV is no longer present on the premises, pigs from a negative source should be introduced a month following the cessation of clinical signs. If the negative pigs do not develop PEDV clinical signs for a 30-day period following introduction to the herd, and a serological test reveals no antibodies to the virus, then PEDV has likely been eliminated from the herd (Geiger and Connor, 2013).

As PEDV elimination is mainly focused on developing immunity in older pigs, high piglet mortality can be expected. Given the economic costs involved in elimination, prevention is still key. To this end, surveillance plays a key role in the early detection of the virus, which facilitates containment and prevents further spread. In Canada, surveillance sampling led to the identification of the first PCR-positive farm in Ontario, which in turn led to further sampling at various locations in the swine production system, and preventive measures at the downstream sites (Kochhar, 2014). The surveillance process, alongside collaborative efforts by federal, provincial, and industry organizations, made the PEDV incursion into Canada less severe than the United States.

1.2.2 Animal Disease Surveillance

Disease monitoring has been defined as the “ongoing efforts directed at assessing the health and disease status of a given population” while disease surveillance is a “more active system and implies that some form of directed action will be taken if the data indicates a disease prevalence or incidence above a certain threshold” (Salman, 2003). Animal disease surveillance objectives include the detection of emerging or re-emerging diseases, declaration of freedom from specific diseases and infections, monitoring of endemic disease to facilitate estimation of
disease frequency, and detection of endemic disease cases (RISKSUR Consortium, 2015).
According to the definition by Salman (2003), disease surveillance systems require a defined
disease monitoring system, disease level thresholds at which action will be taken, as well as
predefined intervention strategies. The data collection process for a disease monitoring system
can either be active (regular recording of cases for a specific disease) or passive (reporting of
suspect cases to health authorities at the discretion of health professionals). A disease control
program (DCP) combines surveillance as well as disease control and intervention strategies in
order to reduce the frequency of a disease over an extended period of time. On the other hand, a
disease eradication program (DEP) extends a DCP by not simply aiming to control the disease,
but rather eliminating it by removing its causal pathogen or agent (Toma et al., 1999).

Christensen (2003) describes a disease control model which begins with passive
monitoring, moves on to active monitoring, and then surveillance for DCP and DEP purposes. In
the passive monitoring phase, no disease is present, and the monitoring system is simply seeking
unusual patterns. When a disease is introduced and begins to spread within the population, the
monitoring system highlights the pattern as a deviation from the norm, and monitoring then
enters the active phase, whereby efforts are made by farmers, veterinarians, laboratory staff, and
researchers to understand and formally identify the disease. At this phase, some farmers and
industry groups may choose to initiate a voluntary DCP, however, non-voluntary DCPs
(administered by governmental organizations) are usually instituted when the disease is a well-
known foreign animal disease (e.g. an OIE-listed disease), a zoonotic disease, or a disease with
significant economic impact. Once DCP efforts indicate low disease prevalence and eradication
is an option, the DCP evolves into a DEP, and efforts are made to detect the last case by testing
the population most likely to be free of the disease. A successful DEP results in a disease-free
population, and efforts now shift to preventing re-introduction of the disease as well as documenting disease freedom. Assuming the disease is not re-introduced, monitoring returns to the top of the model again, where passive monitoring is initiated to detect unusual disease patterns.

The World Organization for Animal Health (OIE) was established in 1924 with the goal of improving animal disease control globally (OIE 1924a, OIE 1924b). Each year, the OIE publishes a list of animal diseases (based on zoonotic potential, economic, and trade impact) and makes this information publicly available on its website and disease notification systems – database interface WAHID (World Animal Health Information Database) and WAHIS (World Animal Health Information System) (OIE 2017a, OIE 2017b). OIE member countries are required to submit immediate notification (within 24 hours) for exceptional epidemiological events, such as: (i) first occurrence or re-occurrence of a listed disease and/or infection; (ii) occurrence of a listed disease in a new aquatic host species; (iii) first occurrence of a new strain of a pathogen of a listed disease; (iv) a sudden and unexpected increase in the distribution, incidence, morbidity or mortality of a listed disease; (v) an emerging disease with significant morbidity or mortality, or zoonotic potential; and (vi) evidence of change in the epidemiology of a listed disease. Weekly follow-up reports to the immediate notification are required to inform of changes to the epidemiological situation, while a final report is required to advise if the disease outbreaks have ended or the situation becomes endemic (Jebara et al., 2017).

Six-monthly reports and annual reports are also required of OIE member countries. The six-monthly report is intended to provide information on the presence or absence of an OIE-listed disease, preventive measures to ensure the disease is not introduced to the country, or control measures if the disease is already present (Jebara et al., 2017), while annual reports
should include information required in the six-monthly reports, as well as data on the country’s animal population, veterinary personnel, human zoonotic cases, national Reference Laboratories, as well as the production of vaccines (Jebara et al., 2017).

In the United States, animal disease surveillance is conducted by the United States Department of Agriculture (USDA), specifically via the Animal and Plant Health Inspection Services (APHIS) division. The APHIS conducts surveillance via its National Animal Health Surveillance System (NAHSS), which is a Veterinary Services initiative designed to integrate existing and future animal health monitoring and surveillance systems (Tomlinson, 2010). There are two main reporting systems within the NAHSS umbrella, namely, the National Animal Health Reporting System (NAHRS) for OIE-listed diseases (USDA, 2017b), and the National Animal Health Monitoring System (NAHMS) for livestock, aquaculture and poultry health management (USDA 2017c). In addition to the NAHRS and NAHMS, the NAHSS also includes national disease surveillance programs for specific animal streams, including bovine, equine, avian, and aquaculture (Tomlinson, 2010). NAHSS surveillance testing is conducted by the National Animal Health Laboratory Network (NAHLN), a coordinated network and partnership of federal, state, and university-associated animal health laboratories (Tomlinson, 2010).

The NAHRS, an initial joint effort of the U.S. Animal Health Association (USAHA), the American Association of Veterinary Laboratory Diagnosticians (AAVLD), and the APHIS, gathers data from chief state animal health officials on the presence of confirmed OIE-listed diseases in the United States (USDA 2017b). The NAHMS combines the efforts of multiple government agencies, producers, academia, and animal health professionals in conducting national studies on animal health (USDA 2017d). NAHMS data are obtained from the following sources, amongst others: (i) diagnostic laboratory disease reporting - from the Veterinary
Diagnostic Laboratory Reporting System (VDLRS)- reports test results from voluntarily participating laboratories across the United States; (ii) sentinel cattle feedlot reporting - voluntary system which provides monthly death loss reporting by veterinarians working in cattle feedlots; (iii) National Agriculture Statistics Service (NASS) Inventory Information - provides data on cattle, sheep, and swine; (iv) National Agriculture Statistics Service (NASS) - Death Loss Studies; (vi) National Poultry Improvement Plan (NPIP) Breeder Pathogen Information; (vii) Somatic Cell Count (SCC) Information from the Diary Herd Improvement Association (DHIA) and the USDAs Agricultural Marketing Service (AMS); and (viii) Food Safety Inspection Service (FSIS) Slaughter Condemnation Data (Wineland and Dargatz, 1998).

In Canada, animal disease surveillance is conducted by the Canadian Food Inspection Agency (CFIA) (CFIA, 2017a). The CFIA has four broad categories for animal diseases, namely, (i) immediately notifiable diseases: diseases exotic to Canada for which there are no control or eradication programs. Diagnostic laboratories are the only entities required to contact the CFIA regarding suspicion or diagnosis of one of these diseases; (ii) reportable diseases: animal owners, veterinarians and laboratories are required to immediately report the presence of an animal that is infected or suspected of being infected with one of these diseases to a CFIA district veterinarian; (iii) annually notifiable diseases: diseases for which Canada must submit an annual report to the World Organization for Animal Health (OIE) indicating their presence within Canada. In general, these diseases are present in Canada, but are not classified as reportable or immediately notifiable; (iv) monitored but not notifiable: diseases of interest that are not currently reportable diseases or notifiable diseases (CFIA, 2017b).

The CFIA’s surveillance efforts are supported by the Canadian Animal Health Surveillance Network (CAHSN), various animal traceability systems (CFIA, 2017a), as well as
the National Aquatic Animal Health Program (NAAHP), which is co-administered with Fisheries and Oceans Canada (CFIA, 2017c). The CAHSN was modeled after the Canadian Network for Public Health Intelligence (CNPHI) (Mukhi et al., 2007), with the objective of providing early detection of animal disease threats, especially foreign and emerging animal diseases. It is a network that links national, provincial, university and private animal health laboratories as well as animal experts and specialists from government and academia. The CAHSN’s initial surveillance focus was on highly pathogenic avian influenza (HPAI), foot and mouth disease (FMD), classical swine fever (CSF), and Newcastle disease (ND), however, CAHSN has also been involved in surveillance for bovine spongiform encephalopathy (BSE) and swine influenza (Kloeze et al., 2010).

CAHSN consists of four components, namely, (i) laboratory diagnostics - the National Centre for Foreign Animal Diseases in Winnipeg Manitoba serves as the OIE reference laboratory and provides standardized methodologies, training, and reagents for Foreign Animal Disease diagnostics in all CAHSN network laboratories; (ii) surveillance - coordination of surveillance activities across provinces, veterinary schools, and the CFIA, achieved through the development of a "Minimum Data Set" which specifies the data elements to be collected, definitions and standards applied to the data elements, as well as appropriate case definitions; (iii) information and technology platform - CAHSN makes use of CNPHI's 'Smart Engine' technology, which transforms incoming data into a common format for better analysis and reporting (Aramini and Mukhi, 2007). The 'Smart Engine' enables daily data transfers from CAHSN network laboratories to the central platform; (iv) governance - as the CFIA and provincial authorities have different disease monitoring mandates based on their respective animal health legislation, an inter-jurisdictional CAHSN Steering Committee was formed to
facilitate information-sharing between network partners, while remaining in compliance with federal and provincial privacy laws (Kloeze et al., 2010).

The CAHSN is a member of the Canadian Animal Health Surveillance System (CAHSS), a non-regulatory network of animal health surveillance networks at the federal and provincial levels, within academia and veterinary organizations, as well as across industry organizations and laboratories (CAHSS, 2017a). Some of its 58+ linked networks and organizations include the Alberta Veterinary Surveillance Network, Quebec's Animal Health Alert and Information Network (RAIZO), the Canadian Wildlife Health Co-operative (CWHC), and the Ontario Animal Health Network (OAHN) (CAHSS, 2017b).

The Ontario Animal Health Network (OAHN), developed in conjunction with the Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA) and the University of Guelph's Animal Health Laboratory (AHL), is comprised of species-specific expert networks (e.g. bovine, poultry, swine, small ruminants, amongst others), with surveillance data obtained from quarterly surveys distributed to veterinarians as well as laboratory data noting the top pathogens/diseases affecting each species (OAHN, 2017).

While surveillance is effective at detecting changes in the health and disease status of a population, recent advances in predictive analytics can provide a more proactive approach to surveillance, whereby data gathered within the framework of a DCP or DEP can be used to forecast future cases. In this sense, predictive methods can serve as a complement to the surveillance toolkit – rather than simply reacting to a disease outbreak, a forecast indicating a potential outbreak or a higher-than-normal increase in cases can assist policy makers in mitigating disease spread by effectively allocating often-limited resources.
1.2.3 Predictive Modelling

1.2.3.1 Overview

Predictive modelling, or forecasting, has been defined as "the process of predicting future events based on foreknowledge acquired through a systematic process or intuition", with data, information, and advanced knowledge as core requirements of the process (Soyiri and Reidpath, 2012; Armstrong, 2001; Lewis et al., 2011). Forecasting methods have wide applications in several fields, including economics and commerce (Armstrong, 2001; Makridakis and Taleb, 2009), sports (Vaughan and Stekler, 2010; Vaughan, 2005), the environment (McMichael et al., 2003; World Meteorological Organization, 2009), and health (Marno et al., 2010; Met-Office, 2017; World Health Organization, 2005; Rogers et al., 2010). Furthermore, health forecasting involves the prediction of "health situations or disease episodes”, forewarning about future events, and serves as a "form of preventive medicine or preventive care aimed at improving health care service provision in populations" (Soyiri and Reidpath, 2012; Met-Office, 2017; Rogers et al., 2010; Van Meijgaard et al., 2010; Sekhri et al., 2010). Health forecasting has been successfully applied to the prediction of emerging infectious diseases, such as the 2014-2015 Ebola epidemic in Liberia (Venkatramanan et al., 2017), as well the prediction of epidemic diseases, such as severe influenza and dengue fever in Taiwan (Li et al., 2016).

Predictive methods have evolved over the years, from classical statistical techniques such as linear and logistic regression, quantile and polynomial regression, Poisson and negative binomial models, autoregressive integrated moving average (ARIMA), seasonal autoregressive integrated moving average (SARIMA) and Holt-Winters exponential smoothing, to machine learning non-parametric methods, such as support vector machines (SVM), artificial neural networks, random forests, classification and regression trees (also known as decision trees), fuzzy rule-based systems, k-nearest neighbor, and Bayesian classification. The nature of the
data, as well as the predictive task at hand, usually dictate the forecasting approach and in some cases multiple approaches (both data and statistically-driven) may be analyzed in order to obtain the best predictive model.

The following sections focus on three machine learning predictive methods – Classification and Regression Trees (CART), random forests, and artificial neural networks – and their applications in human and animal health. Classification and Regression Trees (CART) work by recursively partitioning a dataset into subgroups, with each successive subgroup more homogenous with respect to the response or outcome variable. Each subgroup is formed via a logical if-then condition on values of the explanatory variables (Breiman et al., 1984). The Random Forest method is an ensemble method, in the sense that it “grows” many CART trees. Given a set of input variables and their associated values (i.e. an observation), each tree provides its predicted value for the observation (when the response variable is continuous) or a prediction for the class in which the observation belongs (when the response is categorical). For a continuous response, the final prediction for an observation is the mean of all tree responses, while for a categorical response, the final prediction for an observation is the majority vote across the trees (Breiman, 2001). Artificial Neural Networks (ANN) are mirrored after the structure of the human brain, where processing nodes, similar to biological neurons, are organized into groups of one or more layers. A sequence of input and output layer(s), with hidden layers in-between, are interconnected. The hidden layers generate mapping functions which attempt to explain the relationship between the input variable(s) supplied to the input layer and the output variable(s) supplied to the output layer(s). The interaction between the processing nodes, or neurons, occur along connection paths at different connection strengths called weights (Maren, 1990).
1.2.3.2 Classification and Regression Trees

A CART model was used to identify key weather variables contributing to the incidence of dysentery in a Chinese province (Liu et al., 2017). Prior studies had indicated a correlation between dysentery incidence and weather variables, and as such, the authors set out to identify the actual variables contributing to dysentery incidence. Dysentery was the second most-commonly reported infectious disease in Guangxi Province (Liu et al., 2015). Thus, Binyang County of Guangxi Province was chosen as the study area. Monthly dysentery surveillance data for 2004–2010 were obtained from the Chinese Center for Disease Control and Prevention (CDC), with population data obtained from the sixth census by the National Bureau of Statistics of China, and climate data obtained from the Land Surface Model around the Chinese mainland for 1960–2010.

The initial weather variables selected for investigation were: monthly average precipitation, temperature, relative humidity, wind speed, maximum temperature, minimum temperature, and absolute humidity. Using principal component analysis (a variable reduction method), the variables were reduced to monthly average minimum temperature, precipitation, and relative humidity. The weather variables were then stratified into “low”, “middle”, and “high” levels based on descriptive statistics, specifically, a “low” level was assigned if the corresponding value for the weather variable was between the minimum and 1st quartile, “middle” was assigned for values between the 1st quartile and 3rd quartile, while “high” was assigned for values above the 3rd quartile. The monthly mean incidence was defined as the monthly number of cases of dysentery, divided by the size of the study population, multiplied by 100% to obtain a percentage value. This monthly incidence value was then multiplied by 1000 to obtain the monthly incidence per 1000 people. For the CART model, the explanatory
variables were monthly average minimum temperature, precipitation, and relative humidity, while the response variable was the monthly incidence per 1000 people.

The CART model revealed that dysentery incidence values were lowest when minimum temperature was at the “low” level. With minimum temperature at the “middle” level, relative humidity at ‘low” or “middle” levels, and precipitation at “middle” or “high” levels, dysentery incidence increased significantly. With minimum temperature at the “high” level, and precipitation and relative humidity at “middle” levels, dysentery incidence decreased significantly. The highest incidence was observed when minimum temperature was at a “high” level and relative humidity was also at a “high” level. The second-highest dysentery incidence was observed with minimum temperature at a “high” level, relative humidity at a “middle level” and precipitation at a “high” level. The authors therefore concluded that when minimum temperature was at a “high” level, high incidence of dysentery occurred if relative humidity or precipitation was at a “high” level. High minimum temperatures, high humidity, and high precipitation were identified as the most favorable climate conditions for dysentery transmission. The finding is of public health importance, as it indicates that dysentery intervention efforts should be intensified when the weather is hot and humid and accompanied with heavy rainfall.

In another study, CART analyses were conducted to illustrate the use of the method as an early warning system for emerging infectious diseases (Saegerman et al., 2012). The first case study pertained to the early detection of Bovine Spongiform Encephalopathy (BSE), a neurogenerative disease in cattle with zoonotic potential (Saegerman et al., 2012). With BSE on the decline in many European countries and a lowering of control measures, veterinary reporting of clinically-suspected BSE cattle is the most common approach for detecting sporadic cases of BSE. As such, improvements to clinical diagnosis of BSE are important. To demonstrate the
ability of the CART method to improve such diagnosis, a CART model was built with a dataset of 302 suspected BSE cases (30 BSE-confirmed cases in Belgium before October 2002, and 272 suspected cases that were subsequently determined to be negative). The explanatory variables for the model were age of the animal (in months), year of birth, number of relevant BSE signs noted, and number of clinical signs typical of listeriosis reported, while the response variable was classification into BSE-confirmed or unconfirmed cases. All 30 BSE-confirmed cases were correctly identified, and the model had 100% sensitivity and 85% specificity. Furthermore, using the criteria tested at each CART tree split for the explanatory variables (e.g. 54 < age < 112 months, score BSE signs >= 4, and year of birth >= 1983), veterinarians can more effectively identify affected cows and retrieve them from the food chain.

The second case study pertained to clinical signs associated with Bluetongue (BT), a non-contagious disease of ruminants caused by the bluetongue virus (BTV) (Saegerman et al., 2012). The severity of BT infection depends on multiple factors, including the age and species of the animal, as well as the virulence of the BTV strain involved. Factors contributing to BTV virulence are poorly understood, however, clinical manifestations of the virus have been closely linked to its resultant vascular injuries. The European BTV-8 outbreak was marked by some unusual features, among which was the severity of lesions in affected cattle. To better understand the virulence of the BTV-8 strain, an investigation into its clinical signs was conducted. Forty-one cattle from Belgian and French farms with confirmed BTV-8 were monitored from the onset of clinical signs to describe the disease pattern. On each farm visit, a standardized clinical form (requesting general information as well as the severity of 50+ clinical signs) was filled for each animal by a veterinarian, with a clinical score calculated for every week until the end of clinical signs.
A CART model was then constructed with these data, with explanatory variables as the 50+ clinical signs and the response variable as the clinical score. Epidemiologists then analyzed the CART model to determine the most important clinical signs every week for the first 7 weeks. The results found that the highest clinical scores were recorded within 2 weeks of clinical onset, and the first recorded clinical signs were rather obvious (e.g. conjunctivitis, lesions of nasal mucosa, and nasal discharge). In addition, skin lesions, a drop in milk production, and weight loss appeared as the disease progressed. Such results could be used by veterinarians to improve the detection of BT in cattle, while a similar CART analysis could aid in the early detection of emerging diseases.

1.2.3.3 Random Forest

A study investigated the utility of the random forest model as an aid in the diagnosis of acute appendicitis at a regional hospital in Taiwan (Hsieh et al., 2010). Acute appendicitis (AAP) is a common medical condition that often requires emergency surgery, however, the decision to operate still presents a challenge for surgeons, as high negative appendectomy rates (i.e. high rates of false diagnoses) have been reported in recent years, indicating that most operations were unnecessary. Furthermore, a clinical suspicion of AAP usually requires further tests, such as active observation, clinical scoring systems, abdominal computed tomography (CT), ultrasonography (US), and diagnostic laparoscopy. Diagnostic laparoscopy is an invasive procedure, and while both CT and US are useful tests, a CT exam usually comes at a high cost, while the US exam relies heavily on the knowledge of the operator. The Alvarado score, a common clinical scoring system, is based on eight clinical features (migration of pain, anorexia, nausea/vomiting, right lower quadrant (RLQ) tenderness, rebounding pain, body temperature, white blood cell count, and neutrophil count) and a corresponding feature score. A total of all
feature scores is rated as follows: score <5: appendicitis unlikely; score 5 or 6: appendicitis possible; score 7 or 8: appendicitis likely; score 9 or 10: appendicitis highly likely. The goal of the study was to compare AAP diagnosis from a random forest model and other machine learning methods with AAP diagnosis from the Alvarado score (Hsieh et al., 2010).

The dataset consisted of 180 patients, all of whom had a consultation session with a surgeon for suspected appendicitis between January 2006 and December 2008. 115 patients (91%) were diagnosed with AAP after pathologic examination of tissue samples, and 65 (9%) were diagnosed without AAP. All 115 patients with AAP received an operation, while 11 of the 65 without AAP also received an operation. A review of the clinical characteristics for the 180 patients revealed that those with AAP (115) had a mean ± SD Alvarado score of 6.9 ± 1.6, while those without AAP had a mean ± SD Alvarado score of 4.9 ± 2.0. The overall mean ± SD Alvarado score for all 180 patients was 6.2 ± 2.0.

Sixteen clinical signs often used in diagnosing AAP were initially selected as input variables for the machine learning models, however, these variables were reduced to nine with feature selection methods, specifically consistency subset evaluation (CSE) and the exhaustive search method. The explanatory variables included in the random forest and machine learning models were body temperature, migration of pain, anorexia, rebounding pain, white blood cell count, neutrophil count, diarrhea, progression of pain, and right flank pain, with the response variable set as the diagnosis of acute appendicitis. Patients with an Alvarado score >= 7 (based on the eight clinical features mentioned earlier) were considered positive for appendicitis. 75% of the dataset was used for model training, with the remaining 25% set aside for model testing.

The random forest model had an overall accuracy of 96%, sensitivity of 94%, specificity of 100%, positive predictive value of 100%, and negative predictive value of 87%, while the
Alvarado score had an overall accuracy of 80%, sensitivity of 84%, specificity of 69%, positive predictive value of 87%, and negative predictive value of 64%; the random forest performed better and was significantly more accurate than the Alvarado score. The authors also noted that the ease of model construction was much greater for random forest than the other machine learning models, as they only had to adjust one key parameter, which was the number of trees in the model. Compared to other diagnostic methods, the random forest model was found to be a low-cost and non-invasive aid in the clinical diagnosis of AAP.

In another study on the use of random forest models, Xie et al (2016) investigated the influence of climate change on the distribution of the emerging amphibian fungus, *Batrachochytrium dendrobatidis* (Bd). Bd is the pathogen for chytridiomycosis, an infectious disease which causes anorexia, lethargy, and lack of righting reflex in amphibians (Berger et al. 2005a, Berger et al. 2005b). It has infected approximately 700 amphibian species globally, and is widely believed to be the leading cause of global amphibian population decline (Berger et al., 1998; Daszak et al., 2003). As Bd has no specific amphibian host, it is likely that many more amphibian species will become hosts as the fungus emerges in their regions. Prior studies suggest that the emergence of the fungus may be due to its spatial spread into regions without prior occurrence of Bd, as well as enzootic to epizootic shifts driven by environmental change, specifically climate change. As such, the goal of the study was to examine associations between Bd occurrence at a site and climate and ecological variables.

The dataset included approximately 5000 Bd sampling records as well as 13 temperature and precipitation variables, including mean annual temperature, highest monthly average temperature, lowest monthly average temperature, and annual mean of monthly precipitation. The ecological variables considered were: altitude of a site, whether the site had documented
enigmatic amphibian declines, estimated number of amphibians at the site, and biome
designation at the site. Random forest models were fitted to the data (entire dataset and dataset
by region), and two sets of predictions for Bd occurrence probabilities were generated, the first
based on current climatic conditions, and the other on predicted climatic conditions for the year
2100 by the Hadley Global Environment Model 2–Atmosphere Ocean (HADGEM2—AO), with
simulations from the International Panel on Climate Change (IPCC). Predictions from the model
based on current climatic conditions accurately captured spatial correlation patterns of the
current Bd distribution, an observation which supports climate variables as strongly influencing
the current spread of Bd.

Furthermore, predictions from the model based on future IPCC climate scenarios indicated
an increase in Bd occurrence probability in the Southern Hemisphere (coastal regions such as
Australia and Africa, mountainous terrain of higher elevations such as the Andes in South
America, Ethiopian highlands, and the highlands along the Great Rift Valley in Africa), and a
decrease in Bd occurrence probability in the Northern Hemisphere (areas of high elevation such
as the Kunlun, Qinling, Taihang, and Nanling mountains in China, Hida and Ou mountains in
Japan, the Rockies, Appalachians, and Interior Highlands of the south-central United States in
North America, and the Alps, Kjolen and Ural mountains in Europe). Such projections could aid
in the design of monitoring and conservation efforts to reduce the risk of Bd infection in
vulnerable amphibian species.

1.2.3.4 Artificial Neural Networks (ANN)

Kara et al. (2006) investigated the use of artificial neural networks as an aid in the diagnosis
of optic nerve disease via the analysis of pattern electroretinography (PERG) signals. The
pattern electroretinography (PERG), a retinal response that can provide information about inner
retinal cells and the macula, has been used to monitor the progression of optic nerve diseases such as ocular hypertension, glaucoma, optic neuritis, optic atrophy and amblyopia. However, a key disadvantage of the PERG is that it requires constant monitoring by the ophthalmologist to detect subtle changes. As such, the aim of the study was to design an artificial neural network that would not only simplify diagnosis, but also enable the physician to arrive at a quicker and more confident diagnosis of optic nerve disease.

PERG signals for 256 individuals (composed of 100 healthy and 156 diseased optic neuritis subjects) were obtained from a Tomey Primus 2.5 electrophysiology unit at the Ophthalmology Department of Erciyes University Hospital. The optic nerve disease patient group had an age range of 31 – 75 years (mean age = 52) and involved 61 male and 75 female patients, while the healthy group had 56 male and 44 female patients, with a mean age of 31 years and an age range of 20 to 52 years. All data recorded from the patients were used as inputs to the ANN, a multilayer perceptron (MLP) trained with the Levenberg Marquart (LM) learning algorithm, with the response variable set as classification into a healthy or diseased group. The PERG signals generated 512 input variables which were subsequently reduced to 16 using a feature reduction method. In addition, the training dataset consisted of 50 healthy and 70 optic nerve disease patients (120 observations), while the test dataset was made up of 50 healthy and 86 optic nerve disease patients (136 observations). The resulting ANN model had 96.4% sensitivity, 90.4% specificity and a positive predictive value of 94.2%.

The authors note that the high sensitivity and specificity values of the ANN PERG analysis indicate that the model can be used as a complement to existing ophthalmic electrophysiology devices when diagnosing optic nerve diseases. PERG signals generated from such devices are sometimes fuzzy, and rather than relying on visual inspection alone (which
could lead to a false diagnosis), a physician can employ the ANN as an additional verification measure to confirm the presence of eye diseases.

In another study on the use of artificial neural networks for disease prediction, Leung and Tram (2000) set out to predict shrimp disease occurrence using an artificial neural network. The global shrimp farming industry experienced strong growth in the 1980s mainly due to technological breakthroughs, such as hatchery practices and feed formulation. However, diseases emerged as a major constraint to the sustainability of shrimp aquaculture, with viral diseases reducing shrimp production and slowing the growth of the industry since 1991. Many diseases have been linked to environmental deterioration as well as stresses associated with farm intensification.

In response to the disease problems faced by the Asian shrimp industry, a study was conducted in 1994 by the Asian Development Bank (ADB) and Network of Aquaculture Centres in Asia-Pacific (NACA), with the aim of understanding environmental problems and their economic impacts via a farm-level survey. The 1994 study was the result of a recommendation by a previous ADB–NACA 1990 study which found that diseases of aquatic animals and plants are closely linked with the environment, and as such, environmental issues should be viewed in the broader context of fish farming systems, design, site selection and management. With this observation, the authors set out to predict the occurrence of shrimp diseases based on farm site selection, design, and farm management practices, specifically shrimp disease outbreaks in Vietnam using ADB–NACA farm-level data from 480 Vietnamese shrimp farms.

The dataset consisted of 68 variables (16 continuous, 52 categorical) describing the farm site selection, design, and farm management practices of 480 Vietnamese shrimp farms (86 semi-intensive and 394 extensive farms). The response variable was a classification of farms into
those with a disease outbreak and those without a disease outbreak. The dataset was randomly divided into two sets: a training dataset with 369 observations and a validation dataset with 111 observations. The dataset was fitted to artificial neural network and logistic regression models in order to compare the performance from a machine learning approach with that of a traditional statistical technique. Using stepwise variable selection (both forward and backward), the initial 68 variables were reduced to 6 for inclusion in the final logistic regression model. Specifically, the following six variables were included in the final logistic regression model: POLYCULTURE (yes/no), DRY POND (yes/no), I/D-CANAL (water discharge into intake/drainage canal; yes/no), WATER-SOURCE (estuary/river, direct-from-sea, canal-from-sea, other), SITE-SELECTION (site selection to avoid impacts of other users; yes/no), SILT-DEPOSIT (deposit silt on-farm, yes/no). Two artificial neural networks were constructed, the first using all 68 initial variables, and the second using the six variables included in the logistic regression model.

The classification accuracy of the logistic regression model was 80.18%. The ANN model with the full set of input variables had an accuracy of 86.49%, while the accuracy with six input variables was 73.87%. The results showed that the ANN model using the full set of input variables had better predictive power than the final logistic regression model with six variables. However, the accuracy of the ANN model with six variables was worse than the accuracy of the logistic regression model with the same variables.

The authors note that although ANNs are superior to traditional statistical techniques in detecting complex nonlinear relationships between explanatory and response variables, ANNs are a black box in the sense that while contributing weights for explanatory variables are known, no insights are provided into each variable’s influence on the response variable. As such,
logistic regression appeared to be the preferred choice for exploring risk factors for shrimp disease due to the non-parametric and black-box nature of the ANN.

Furthermore, logistic regression provided meaningful parameter estimates with insightful policy implications. For example, farms which practiced polyculture, dry ponds, and conducted careful site selection had a smaller chance of a disease outbreak, while farms that deposited silt on the farm, and discharged water into intake and drainage canals had a much higher chance of a disease outbreak.

1.3 Objectives

As previously noted, prevention is still the best course of action against the high mortality and economic costs associated with Porcine Epidemic Diarrhea (PEDV) and Porcine Deltacoronavirus (PDCoV). However, outbreaks will undoubtedly occur, and in such situations, it is helpful to have a quantitative assessment of the outbreak, specifically the measures of disease frequency. When processed and analyzed with machine learning predictive techniques, the outbreak data and measures of disease frequency can yield insights into current disease patterns, as well as short and long-term forecasts for the disease. Such short and long-term forecasts can ensure that often-limited disease control resources are allocated appropriately.

In Chapter 2, PEDV and PDCoV outbreak data for the province of Ontario are retrieved from an industry database, and analyzed with statistical software in order to generate weekly incidence and prevalence counts, as well as weekly incidence and prevalence proportions. This chapter also provides a brief overview of the PEDV and PDCoV surveillance components in Ontario. In Chapter 3, three machine learning methods – Classification and Regression Trees (CART), Random Forest, and Artificial Neural Networks – are applied to measures of disease frequency from Chapter 2 in order to generate weekly long-term and short-term forecasts for
both diseases. Chapter 4 provides a brief summary of related work, accomplishments of the thesis, and suggestions for future research.
References


Rodak, L., B. Smid, L. Valicek, R. Smitalova, and Z. Nevorankova, 2004: ELISA detection of group a rotavirus, transmissible gastroenteritis virus and porcine epidemic diarrhoea virus in


CHAPTER TWO – Herd-Level Prevalence and Incidence of Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) in Swine Herds in Ontario, Canada

2.0 Summary

Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) were first identified in Canada in 2014. Surveillance efforts have been instrumental in controlling both diseases. In this study, we provide an overview of surveillance components for the two diseases in Ontario (Canada), as well as PEDV and PDCoV incidence and prevalence measures. Swine herds located in the province of Ontario, of any type, whose owners agreed to participate in a voluntary industry-led disease control program (DCP), and with associated diagnostic or epidemiological information about the two swine coronaviruses, were eligible to be included for calculation of disease frequency at the provincial level. PEDV and PDCoV data stored in the industry DCP database were imported into the R statistical software and analyzed to produce weekly frequency of incidence counts and prevalence counts, in addition to yearly herd-level incidence risk and prevalence between 2014 and 2016. The yearly herd-level incidence risk for PEDV, based on industry data, was 13.5%, 3.0%, and 1.4% (95% CI: 11.1-16.2, 2.0-4.2, 0.8-2.3) while the yearly herd-level incidence risk for PDCoV was 1.1%, 0.3%, and 0.1% (95% CI: 0.5-2.2, 0.1-0.9, 0.0-0.5), for 2014, 2015 and 2016, respectively. Herd-level prevalence estimates for PEDV in the last week of 2014, 2015, and 2016 were 4.4%, 2.3%, and 1.4% respectively (95% CI: 3.1-6.0, 1.5-3.3, 0.8-2.2), while herd-level prevalence estimates for PDCoV in the last week of 2014, 2015, and 2016 were 0.5%, 0.2%, and 0.2%, respectively (95% CI: 0.1-1.2, 0.0-0.6, 0.0-0.6). Collectively, our results point to low and decreasing incidence risk and prevalence for PEDV and PDCoV in Ontario, making both diseases possible candidates for disease elimination.
at the provincial level.
2.1 Introduction

Porcine Epidemic Diarrhea (PED) was first described in England in 1971, and its causative agent, Porcine Epidemic Diarrhea Virus (PEDV), was identified in 1978 (Chen et al., 2014). PEDV causes anorexia, vomiting, diarrhea, and dehydration in pigs, resulting in near 100% mortality for piglets during the first few days of life (Hill et al., 2014) and low mortality in older pigs. The virus spreads via the fecal-oral route, either through direct contact with an infected pig or through indirect contact with contaminated fomites. Widespread outbreaks were reported in Europe during the 1970s and 1990s, while epidemics in Asia have caused significant disruption to Asian pig production since 2008 (Williamson et al., 2013). Porcine Deltacoronavirus (PDCoV, also known as Swine Deltacoronavirus, SDCV) was first identified in Hong Kong in 2012. The transmission modes and clinical signs due to PDCoV infections are similar to PEDV, however, the mortality rate is generally lower after PDCoV infections (Carvajal et al., 2015).

PEDV emerged in North America in May 2013 while PDCoV was first confirmed in February 2014, both in the United States (Chen et al., 2014; Ma et al., 2015). These novel viruses rapidly disseminated throughout the US swine population, resulting in the mortality of an estimated seven million animals by May 2014 (Jung & Saif, 2015). Due to the fecal-oral route of transmission, the infection spread through various mechanisms, including contaminated transportation vehicles (Lowe et al., 2014). In Canada, PEDV emerged in January 2014 when a swine herd in Ontario tested positive for the virus (Kochar, 2014). Imported spray-dried porcine plasma contaminated with PEDV was the likely pathway of introduction, as established through descriptive studies (Pasma et al., 2016), analytical epidemiological studies (Aubry et al., 2017; O’Sullivan, 2015), and experimental investigations (Pasick et al., 2014). By July 2014, only 62
cases of PEDV had been reported in Ontario and the outbreak was largely under control (Pasma et al., 2016).

Despite considerable impact on animal health in completely susceptible populations, the diseases caused by the two emerging porcine coronaviruses were not notifiable globally, according to the World Organization for Animal Health (OIE, 2017), and were not federally reportable or notifiable in Canada (Canadian Food Inspection Agency, 2014). At the provincial level, animal health regulations enabled consideration of PEDV and PDCoV as emerging hazards (Government of Ontario, 2009), which allowed a measured and appropriate response to the outbreak by the Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) via close collaboration with swine industry organizations.

Initial emergence of the two novel porcine coronaviruses was followed by their successful elimination from several initial case farms (Misener, 2015). The high rate of successful PEDV elimination from individual herds, and effective measures that seemed to have minimized widespread viral dissemination, resulted in the current position of industry organizations that both infectious agents can and should be eliminated at the provincial level (Ontario Swine Health Advisory Board, 2017).

A disease control program (DCP) involves disease monitoring, surveillance, intervention, and control strategies (Salman, 2003). The DCP considered here has been voluntary in nature as defined elsewhere (Christensen, 2003). Furthermore, an important component of any disease control program is measuring trends in incidence and prevalence, particularly when the disease of interest moves into the phase of possible elimination (Salman, 2003). With the infrastructure built for management of endemic diseases in Ontario, the data to support estimation of disease trends are available. Thus, the primary objective of this study is to estimate herd-level incidence
and prevalence measures for PEDV and PDCoV in swine herds in Ontario (Canada) between January 2014 and December 2016, based on industry data. The secondary objective is to describe relevant surveillance components that were used for identification of new PEDV cases.

2.2 Materials and Methods

Data sources

The Ontario Swine Health Advisory Board (OSHAB) maintains a database which contains premises information and PEDV/PDCoV herd status for producers enrolled in its voluntary regional disease control programs. The program, and a database originally designed for management of porcine reproductive and respiratory syndrome virus (PRRSV) (Arruda et al., 2015), was expanded to PEDV and PDCoV after their emergence. The data relevant for this work included unique identifiers, herd type, date of herd enrollment into the database, PEDV and PDCoV status of individual premises, and the date that individual premises changed their PEDV and PDCoV status. For inclusion into the study, swine herds could be of any herd type, but had to meet the following criteria: (i) be located in the province of Ontario, (ii) participate in the voluntary industry-led disease control program, and (iii) have diagnostic or epidemiological information about the infection status of porcine epidemic diarrhea virus (PEDV) or porcine Deltacoronavirus (PDCoV). The industry organization (i.e. OSHAB) provided relevant data to researchers for calculation of disease frequencies under a separate data transfer agreement.

Due to the voluntary nature of the DCP, the enrollment of herds into the database has been an ongoing process. This could have resulted in the date of enrollment being later than the date the infection was originally detected in a specific herd. In rare instances where a herd’s enrollment date wasn’t specified, or occurred later than the first reported case of PEDV or PDCoV for that herd, the disease status date was entered as the herd enrollment date for the
purposes of this report. This was done so that herd-level prevalence on a weekly basis could be properly calculated. An Open Database Connectivity (ODBC) connection to this database was established, and relevant tables were imported into R (R Core Team, 2016) using the `rodbc` library (Ripley & Lasley, 2016).

**Herd-level PEDV and PDCoV infection status definition**

The industry database maintained by OSHAB traces infection status of participating individual premises over time, which includes detection of infection, as well as any subsequent infection, elimination, or other change in infection status with the associated dates. In this database, premises can have four possible values for their infection status, as defined below: (i) **confirmed positive** – premises with herds that tested positive for PEDV/PDCoV based on laboratory test results from the Animal Health Laboratory (AHL) at the University of Guelph (Guelph, ON, Canada). For a premise to be classified as confirmed positive, it had to have an associated diagnostic submission number that includes at least one positive test for PEDV/PDCoV using RT-PCR, regardless of the number of specimens that were submitted. An AHL reference number (also known as a G Number) was not available for some confirmed positive cases in the database, as the attending veterinarian obtained test results, but did not provide OSHAB with the AHL reference number; (ii) **presumed positive** – premises which housed animals that were moved from positive sites at a prior stage in the production system (i.e. defined as positive due to pig movement). This information was obtained from attending veterinarians based on their knowledge of pig flow and movement, and was not based on diagnostic testing conducted on the premises of interest. For premises to be classified as presumed positive, the herd veterinarian simply needed to indicate that a specific site received pigs from PEDV/PDCoV positive sites; (iii) **presumed negative** – previously positive premises,
either confirmed or presumed, that were tested using PCR tests according to industry guidelines and had all test results negative. The sampling requirements for declaring premises to be PEDV/PDCoV presumed negative varied based on the combination of herd type and the type of animal flow (i.e. all in-all out by barn, or continuous flow nursery and finisher herds). Complete criteria were, at the time of publishing, available on the website of the industry organization (Ontario Swine Health Advisory Board, 2015). Briefly, sampling strategy for farrow-wean, nursery, and finisher sites aimed to detect prevalence of virus-positive animals of at least 10%, with expected herd sensitivity of 95%, assumed test sensitivity of 98%, test specificity of 100%. In instances where the sampling material was oral fluid collected through cotton ropes, the assumption was that five pigs contributed oral fluids to one rope, and such fluids were considered a pooled sample. For farrow-finish or farrow-feeder sow sites, the same assumptions were made, except that the sampling strategy was required to detect prevalence of 5% with 95% herd sensitivity. Because of alternative strategies, the required sample size varied but a minimum sample size was four oral fluids. In addition, the testing time in sow herds was prescribed to be a minimum of ten weeks post-infection, and was required to be repeated three times in the case of farrow-wean (FW) sites, or two times in the case of farrow-finish (FF) or farrow-feeder (FG) sites. The recommended specimen type could be swab, Swiffer (for covering larger areas in a pen), or oral fluids, depending on the target age group. In FW herds, individual farrowing crates were the target population for each individual sampling occasion, in particular if diarrhea was evident. The minimum recommendation for one sampling occasion in FW herds was to sample four Swiffer samples, at least eight farrowing crates per one sample. Alternatively, individual swabs of 30 farrowing crates were deemed as acceptable sample after pooling 5:1. In FG farms, the recommended sample type was oral fluid, with recommendation to collect 12 oral fluids from
nursery pigs. Similarly, in FF farms, six oral fluids were recommended for collection from nursery pigs, and six for collection from finisher pigs. For all-in/all-out nursery and finisher farms, the recommendation was to sample six oral fluids, with added requirement that these herds should be supplied from sow herds with a confirmed negative status. All testing has been assumed to be performed using RT-PCR tests. Full description is available elsewhere (Ontario Swine Health Advisory Board, 2015); (iv) confirmed negative – premises which have had no clinical signs or diagnostic evidence of PEDV/PDCoV for at least 6 months after the presumed negative status date. In addition, herds that were part of the Ontario voluntary DCP but were not tested for emerging porcine coronaviruses - due to lack of clinical or other types of diagnostic or epidemiological triggers - had assigned status of NA (not available).

**Weekly measures of disease frequency**

Weekly time series were then obtained for the three sets of statistics for both viruses: a) number of premises with specific infection status in each week, b) number of new positive and new negative premises, and c) number of premises in the database.

For each week where a herd’s PEDV/PDCoV status was not reported, the status was set to the last-reported status using the zoo package (Zeileis & Grothendieck, 2005). For example, if a status is not reported for the current week, and ‘confirmed positive’ was reported for the prior week, then the current week’s status is ‘confirmed positive’. The individual premises data were then aggregated to counts of premises on a weekly basis. Based on the former time series, prevalence area plots were generated, providing a visual assessment of 'confirmed positive', 'presumed positive', 'presumed negative', and 'confirmed negative' herds over time. The prevalence numerator was the positive herd count (sum of confirmed positive and presumed positive herds) for a specific week, while the denominator was the herd count in the premises.
table for that week (calculated previously).


Any status change leading to new presumed or confirmed positive status (i.e. status changes 1, 2, 7, 8, 9, and 10) were classified as new positive. Similarly, status changes 3, 4, 5, and 6 were classified as new negative. The number of new positives and new negatives were then aggregated to the weekly level throughout the study period. The latter time series were then used to construct a chart of cumulative incidence counts for each year, and epidemic curves were constructed for positive herds and herds which became negative. In addition, for each week, the incidence risk was calculated by dividing the number of cases that occurred in a specific week, by the number of herds that were eligible to become cases at the beginning of the week.

**Yearly measures of disease frequency**

Yearly herd-level incidence risk and prevalence at the end of the year was calculated in the following manner: the numerator for the yearly incidence risk calculation was defined as the cumulative number of incident cases at the end of each year, while the denominator was defined as: the number of herds in the DCP at the end of each year, minus the number of confirmed positive and presumed positive herds at the beginning of the specified year, minus half the number of herd additions to the database during the year (the number of additions was treated as
the number of withdrawals so this inherently open population could be converted to a closed population, and equations for calculation of risk used) (Dohoo et al., 2003):

\[
\text{Yearly Incidence Risk (} R \text{)} = \frac{N}{PAR}
\]

where \( N = \text{cumulative incident cases at yearend} \)

and \( PAR (\text{Population At Risk}) = \)

\( (\text{number of herds in DCP at yearend}) \)

\[ - (\text{positive herds at beginning of year}) \]

\[ - \left( \frac{1}{2} \times \text{number of additions to DCP during the year} \right) \]

Eq. 1 (adapted from Dohoo et al., 2003)

For yearly herd-level prevalence, the numerator was defined as the total number of confirmed positive and presumed positive herds at the end of the year, while the denominator was defined as the number of herds in the DCP at the end of each year. Both the yearly incidence risk and prevalence, originally calculated as proportions, were then multiplied by 100 to get the percentage values, with exact 95% confidence intervals obtained from the \textit{binom.test} function in R.

For the yearly incidence rate, we calculated total number of herd-years-at-risk for each year from the number of herds under risk in each week, and used this as a denominator. The incidence rate was then expressed as number of cases per one herd-year. Exact 95% confidence intervals on the incidence rate were obtained via the \textit{poisson.test} function in R.

**Surveillance components for PEDV**

Three surveillance components for identification of new cases were considered in this study. The cumulative number of presumed and confirmed cases between January 2014 and December 2016 maintained in the OSHAB voluntary DCP database represents surveillance
component 1. Surveillance component 2 is the official count of new cases maintained by OMAFRA and publicly available when this report was written (Ontario Pork, 2017). OMAFRA does not monitor PDCoV, and defines a PEDV case as a herd which tests positive based on AHL’s RT-PCR test (Pasma et al., 2016), and is the first such case in the production system. In other words, the PEDV case count pertains to primary cases only; subsequent secondary cases due to animal movement in the production system are not included in OMAFRA’s reporting, although they could be confirmed as PEDV positive in the diagnostic laboratory. Surveillance component 3 represents cumulative number of positive submissions, due to any reason, maintained by the Animal Health Laboratory (AHL, University of Guelph), the largest diagnostic laboratory in Ontario for testing livestock diseases. The overlap between the three surveillance components was calculated on the basis of demographic data.

2.3 Results
Surveillance components for PEDV

Figure 2.1 depicts three surveillance components considered in this study and their overlap. Briefly, during the 3-year period (January 2014 – December 2016), AHL reported a total of 974 positive submissions for PEDV (Fairles, personal communication 2017). Of the 974 positive submissions, a cumulative total of 118 cases are reported in OMAFRA and OSHAB as new cases. The remaining 856 positive submissions were not part of any surveillance component considered in this study. Of the 118 cases, 99 are captured by OMAFRA as primary PEDV cases, however, 70 of these 99 primary PEDV cases are also reported in OSHAB’s database, leaving 29 primary PEDV cases (99 – 70) reported in OMAFRA but not OSHAB. The remaining 19 cases (974 – 856 – 99) are non-primary PEDV positive cases (i.e. not the first outbreak in the associated production system), hence they are captured in AHL and reported in OSHAB’s voluntary DCP database, but not OMAFRA. Lastly, 52 PEDV positive cases are
strictly in OSHAB and are not reported in any other surveillance component. It should be noted that these 52 cases are ‘presumed positive’ due to pig movement, and as such have no supporting AHL test result, an attribute which also makes them ineligible for OMAFRA reporting. The OSHAB total for PEDV cases is 52 + 19 + 70 = 141, which is also the cumulative number of incident cases over the 3-year period.

**Yearly incidence and prevalence measures**

Table 2.1 contains number of new infected premises for the two viruses in each year based on the industry’s voluntary DCP database, number of positive cases in the last week of each year, as well as incidence risk and prevalence estimates. In addition, 95% confidence intervals are provided, recognizing that this surveillance component was a census of herds participating in the voluntary DCP, but does not contain all herds from the source population. In brief, annual incidence risk was the highest in 2014 for PEDV (13.49%) and PDCoV (1.14%), and the lowest in 2016 for PEDV (1.42%) and PDCoV (0.08%). The end of year prevalence for PEDV ranged between 4.4% and 1.4% in 2014 and 2016 respectively. A similar trend can be seen for PDCoV (Table 2.1). Furthermore, similar values have been observed for incidence rates in each year (Table 2.1).

**Weekly incidence and prevalence measures**

The cumulative number of new PEDV and PDCoV cases per week for the three years is provided in Figures 2.2 and 2.3, respectively. Notably, there were many new PEDV cases detected in the winter of 2014 when the disease was first introduced to Ontario and Canada.

Figure 2.4 depicts the number of herds categorized into one of the four groups with respect to PEDV infection. From this figure, it could be observed that PEDV had a rapid increase in the number of prevalent cases until March 2014, which then peaked in the summer of the same
year, only to be followed by a decline. Figure 2.5 displays the weekly development of infection status as a proportion. Note that in the latter figure, the denominator changed on a weekly basis as the number of herds in the voluntary DCP changed. Figures 2.6 and 2.7 display the number of premises in a distinct PDCoV status as a count (Figure 2.6) and proportion (Figure 2.7). The weekly incidence risk for PEDV and PDCoV in 2014 - 2016 are provided in the supplementary material (Figure S.2.1 and Figure S.2.2 respectively).

2.4 Discussion

Emerging infections considered in this work are examples of production-limiting diseases that are not federally reportable (Canadian Food Inspection Agency, 2017) or listed in an OIE list of diseases (World Organization for Animal Health, 2017). Consequently, resources available through response by the regulatory veterinary authorities could be limited. In the case of the Ontario outbreak, the regulatory framework existed at the provincial level because PEDV was considered an emerging hazard under the provincial Animal Health Act (Government of Ontario, 2009) and because of suspicion of feed involvement, which is under the jurisdiction of national veterinary authorities. The response to the outbreak was collaborative in nature and is best depicted by the evaluation of surveillance components for PEDV.

The OMAFRA and industry approaches were different with respect to surveillance coverage, case definition, and follow-up time. OMAFRA’s surveillance system was mostly concerned about identification of new cases and their investigation in the early phase of the outbreak. It contained a census of all PEDV cases, and all herds from the source population were eligible to be listed as a case, at least during the initial phase of the outbreak when PEDV could be considered an emerging hazard. However, only the first case in a given production system was counted as a case and secondary cases due to planned animal movement to other premises...
were not counted as additional cases. For example, if PEDV was detected in a sow herd in a multi-site production system, the sow herd would be counted as a case, but not the nursery or finisher sites supplied from the sow herd. For this surveillance system, all cases had to have diagnostic confirmation through laboratory submissions. For this reason, the OMAFRA’s PEDV surveillance system is completely nested within the diagnostic laboratory. It is notable that ~17.1% of PEDV cases were identified through this surveillance system only, which is reflective of the reportable nature (provincially) of the emerging hazard.

The OSHAB surveillance system was concerned with identification of new cases, reducing dissemination of infection through animal movement, and following the infection status over time for the purposes of monitoring disease trends over time. As a voluntary program, the source population consisted of premises that volunteered to participate. At the end of 2016, the number of sites in the OSHAB database was 1255, which represents approximately 49% of 2,556 active swine sites in Ontario (Brisson, 2014). Thus, the OSHAB surveillance coverage was lower than for the OMAFRA surveillance, which had 100% coverage since by law, any PEDV-infected herd must be reported to OMAFRA, at least during the phase when the hazard was considered as emerging.

The OSHAB case definition also included secondary sites that were confirmed positive due to animal movement and had laboratory confirmation, as well as secondary sites where movement of PEDV-positive animals occurred although testing was not done. This approach allowed identification of ~41.8% of cases that would otherwise be unaccounted for in the OMAFRA system (n=71), which included 30.6% of total cases (n=52) that were not confirmed through diagnostic testing (Figure 2.1). Given the purpose of the industry’s surveillance system, recording infection in the secondary sites was a logical choice since they could still contribute to
disease dissemination through multiple pathways. Such premises had to be declared as presumed or confirmed negative, based on diagnostic testing that met pre-determined criteria conducted on the premises.

The Venn diagram of surveillance components (Figure 2.1) also shows a large volume of positive test results for PEDV in a laboratory that did not contribute to the identification of new cases. Such testing serves different purposes, and likely contain additional surveillance components which are not considered here. An example is testing trailers in abattoirs for the purposes of case identification through risk-based approaches. Surveillance systems for emerging diseases could benefit from a clearer definition of such surveillance components, so that diagnostic tests could be easily aggregated into appropriate streams.

The collaborative efforts at controlling PEDV is most clearly seen in the incidence counts for 2014. As expected at the onset of any outbreak, particularly due to massive contamination through a common source, there was a rapid climb in the number of infected farms. However, this outbreak was brought under control relatively quickly. By May 2014, the worst of the outbreak was over, and there were only single digit weekly infections (no more than 5) for the remainder of the year. In addition, the initially positive herds which contributed to the rapid climb in infections between January and May 2014 became negative within a short period, and contributed to the rapid increase in negative herds between June and December 2014.

The incidence counts for 2015 reveal a more proactive approach to OSHAB’s PEDV surveillance and disease control. Not only were there substantially fewer infections (infections were only reported for 15 weeks out of the year, and there were at most 5 infections per week), the number of negative herds also kept up with the number of positive herds (for 9 of the 15 weeks where infections were reported, there were also negative herds). Overall, the number of
negative herds in 2015 outpaced the number of positive herds, indicating proactive steps to minimize the number of new infections while aggressively taking steps to eradicate existing ones. This stands in stark contrast to 2014, where disease eradication efforts only seemed to take on momentum towards the latter half of the year.

The trend in incidence counts from 2015 continued into 2016, with lower numbers for positive herds and the number of negative herds outpacing positive ones. However, in contrast to 2015, roughly 75% of new infections occurred in the first quarter of the year (January – March), while 80% of new negative herds were reported in the last quarter of the year (late September – December). The cluster of negative herds at the end of the year suggests that with new infections under control, emphasis was placed on the management of prevalent cases, taking into consideration lower viral transmission during warmer periods of the year.

The prevalence plots reveal ebbs and inclines in PEDV prevalence, in what appears to be a cyclical pattern: a peak in July 2014, a low in January 2015, a peak in July 2015, a low in October 2015, and a peak in February 2016 (Figure 2.4). The numbers were relatively constant from late February – July 2016, declined in late July, and slowly declined until the end of the year. Given the prevalence fluctuations between 1% - 2% for September 2015 - December 2016 (Figure 2.5), and the repeated patterns, there’s a danger that PED may become an endemic disease with low prevalence and limited pockets of infection in Ontario. As such, further investigation and action is needed to ensure that the peaks are replaced by a steady decline.

As the same disease prevention and control measures for PEDV also apply to PDCoV (Ontario Ministry of Agriculture Food and Rural Affairs, 2014), there were few reported infections for PDCoV (essentially, controlling for PEDV also controls for PDCoV). It is therefore not surprising that the incidence count patterns are similar: in 2014, there were more
PDCoV positive herds than negative herds, while in 2015 there were more negative herds than positive ones. However, in 2016 there was just one new PDCoV infection, with a single herd also becoming negative, an indication that if current disease control efforts continue, there may be no new infections in subsequent years.

The prevalence figures for PDCoV are quite different from PEDV – rather than a cyclical pattern as previously observed, the numbers are more erratic, with sudden peaks followed by constants over prolonged periods, and then declines. The relatively static and low prevalence (0.1% - 0.2%) observed between July 2015 and December 2016 suggest that PDCoV is a candidate for disease elimination. There is also a possibility that PDCoV cases are underreported by the industry, either because of potentially lower clinical impact in swine herds, or because of perceived lower importance than PEDV. Nonetheless, when any of the three porcine coronaviruses (PEDV, PDCoV, TGEV - transmissible gastroenteritis virus) is suspected in a herd and diagnostic material is submitted to the AHL, the diagnostic testing is automatically conducted for all three viruses.

Some limitations of the current study include the fact that PEDV and PDCoV cases are actively pursued for inclusion into the voluntary DCP by industry organizations. Such strategy is likely to result in estimates of incidence and prevalence measures that are higher than in the source population. Because of the inherently open nature of the voluntary DCP, we had to modify the formulae for calculation of incidence risk. It could be argued that presumed cases are not diagnostically confirmed, and are therefore subject to misclassification. However, the reality of a voluntary DCP for production-limiting diseases is that resources to conduct large-scale testing are scarce, and as such, need to be carefully deployed. Furthermore, in order to confirm premises as presumed negative, diagnostic testing to confirm absence of infection at the design
prevalence level is still required. Also, the criteria to declare confirmed negative status is arguably open ended, and could be further improved.

In conclusion, this study provides estimates of incidence and prevalence measures in Ontario based on industry data collected through voluntary disease control programs. The data suggest that annual incidence risk and prevalence estimates are low and have been steadily decreasing between 2014 and 2016 for PEDV and PDCoV. Current estimates of disease frequency support planning of disease elimination at the provincial level, but more information should be available about factors that led to time to elimination in individual herds. In addition, our evaluation of surveillance components indicate that the two surveillance components were complementary, and focused on different aspects of surveillance. OMAFRA’s surveillance was mostly focused on identification of primary cases aimed at quick disease investigations and traceability in the face of the outbreak, whereas OSHAB’s surveillance component has the added benefit of having sufficient data that allows long term evaluation of disease trends, long-term disease management, and tracing disease status of individual herds over time. OSHAB’s voluntary DCP database also provides a good tool for calculating weekly prevalence and incidence measures, which is a valuable statistic for producers and animal health experts during all phases of disease outbreak and control.

2.5 Acknowledgements

Funding for this study was provided by the NSERC Discovery Grants Program. Authors are grateful to participating industry organizations and producers for providing access to data.
References


O'Sullivan, T. (2015, June). Investigation of Factors That Led to Emergence of PEDV through Feed During Early Phase of Canadian Outbreak. Paper presented at the 7th International Symposium on Emerging and Re-Emerging Pig Diseases, Kyoto, Japan


Table 2.1: Herd-level incidence risk and rate of two novel porcine coronaviruses (PEDV and PDCoV) in Ontario swine herds between 2014 and 2016, and estimated prevalence of positive cases at the end of each year based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)

<table>
<thead>
<tr>
<th>Year</th>
<th>Cumulative number of new cases</th>
<th>Incidence Risk (%)</th>
<th>95% CI (%)</th>
<th>Incidence Rate (cases per herd-year)</th>
<th>95% CI (cases per herd-year)</th>
<th>Number of cases at year-end</th>
<th>Prevalence at year-end (%)</th>
<th>95% CI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Porcine Epidemic Diarrhea Virus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>95</td>
<td>13.49</td>
<td>(11.06 - 16.24)</td>
<td>0.14</td>
<td>(0.12 – 0.18)</td>
<td>36</td>
<td>4.36</td>
<td>(3.07 – 5.99)</td>
</tr>
<tr>
<td>2015</td>
<td>29</td>
<td>2.97</td>
<td>(2.00 – 4.24)</td>
<td>0.03</td>
<td>(0.02 – 0.05)</td>
<td>27</td>
<td>2.25</td>
<td>(1.49 – 3.26)</td>
</tr>
<tr>
<td>2016</td>
<td>17</td>
<td>1.42</td>
<td>(0.83 – 2.26)</td>
<td>0.02</td>
<td>(0.01 – 0.03)</td>
<td>17</td>
<td>1.35</td>
<td>(0.79 – 2.16)</td>
</tr>
<tr>
<td></td>
<td>Porcine Deltacoronavirus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>8</td>
<td>1.14</td>
<td>(0.49 – 2.23)</td>
<td>0.011</td>
<td>(0.005 – 0.022)</td>
<td>4</td>
<td>0.48</td>
<td>(0.13 – 1.24)</td>
</tr>
<tr>
<td>2015</td>
<td>3</td>
<td>0.30</td>
<td>(0.06 – 0.87)</td>
<td>0.003</td>
<td>(0.001 – 0.009)</td>
<td>2</td>
<td>0.17</td>
<td>(0.02 – 0.60)</td>
</tr>
<tr>
<td>2016</td>
<td>1</td>
<td>0.08</td>
<td>(0.00 – 0.45)</td>
<td>0.001</td>
<td>(0.000 – 0.005)</td>
<td>2</td>
<td>0.16</td>
<td>(0.02 – 0.57)</td>
</tr>
</tbody>
</table>
Figure 2.1: Venn diagram representing overlap among the three surveillance components for identification of new cases of the Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) between 2014 and 2016 – The Animal Health Laboratory (AHL), Ontario Ministry of Agriculture Food and Rural Affairs (OMAFRA), and the Ontario Swine Health Advisory Board (OSHAB). The size of individual circles should be proportional to the number of positive cases recorded by a specific surveillance component. The overlapping areas should represent cases that were identified by as many surveillance components as are being overlapped. The numbers listed in a specific area represent number of cases that are identified by a specific surveillance component, or combination thereof.
Figure 2.2: Cumulative weekly confirmed and presumed positive cases of Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)
Figure 2.3: Cumulative weekly confirmed and presumed positive cases for Porcine Deltacoronavirus (PDCoV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)
Figure 2.4: Number of Herds with Confirmed Positive, Presumed Positive, Presumed Negative, or Confirmed Negative status for Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)
Figure 2.5: Percentage of Herds with Confirmed Positive, Presumed Positive, Presumed Negative, or Confirmed Negative status for Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)
Figure 2.6: Number of Herds with Confirmed Positive, Presumed Positive, Presumed Negative, or Confirmed Negative status for Porcine Deltacoronavirus (PDCoV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)
Figure 2.7: Percentage of Herds with Confirmed Positive, Presumed Positive, Presumed Negative, or Confirmed Negative status for Porcine Deltacoronavirus (PDCoV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)
Figure S.2.1: Weekly incidence risk for Porcine Epidemic Diarrhea Virus (PEDV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)
Figure S.2.2: Weekly incidence risk for Porcine Deltacoronavirus (PDCoV) in Ontario (Canada) for 2014 – 2016, based on data provided in the Ontario Swine Health Advisory Board (OSHAB) Disease Control Program (DCP) database (average number of herds for 2014 - 2016 = 1093)
CHAPTER THREE – Forecasting Herd-Level Porcine Epidemic Diarrhea (PED) Incident Cases and Trends in Ontario (Canada)

3.0 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 not only in identifying the initial case but also in minimizing the spread of the virus to other farms in the province. 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 and prevalence data, and in so doing determine their accuracy for predicting weekly PEDV counts (short-term forecasting) and future PEDV trends (long-term forecasting). A secondary objective was to identify weather variables with a strong influence on short-term and long-term PEDV forecasts. 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. Random forest, artificial neural network, and classification and regression tree models were fitted to the dataset, with variations in the number of explanatory variables. 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 networks, and classification and regression tree models respectively. The random forest models provided the best prediction for both short and long-term forecasting, while low and average temperatures were strongly associated with weekly PEDV counts and future PEDV trends.

Keywords: porcine epidemic diarrhea, disease surveillance, disease forecasting, random forest, classification and regression trees, artificial neural networks.
3.1 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 is of high quality, and was successfully utilized to obtain PED incidence and prevalence estimates for the province (Ajayi, 2017). With limited surveillance and disease control resources, long-term and short-term predictions for PED 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 & 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.

Therefore, the primary objective of this study was to apply machine learning predictive methods, specifically random forest, classification and regression trees, and artificial neural networks, to the PED dataset to determine the best method for short-term forecasting (weekly
PEDV counts) and long-term forecasting (future PEDV trends). The secondary objective was to determine which environmental variables are highly ranked as determinants for weekly PEDV counts and future PEDV trends.

3.2 Methods
3.2.1 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, 2017). Briefly, individual herds could be classified as: confirmed positive, presumed positive, confirmed negative, presumed negative, and undetermined (which in this database equates to negative due to the emerging nature of the disease and reporting requirements in the early phase of the epidemic). These statuses were then processed to obtain weekly measures of disease frequency, including number of new cases (i.e. the outcome for this study), herd-level prevalence (expressed as a percentage), and number of infectious sows (expressed as raw counts).

Temperature and humidity values representative of swine locations in Ontario were obtained 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; 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.

The daily weather data for each county was then imported into R and a singular data frame 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 (e.g. 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) on\text{ontario\textunderscore hightemp} – highest temperature reading in Ontario for the day, (ii) on\text{ontario\textunderscore avgtemp} – average temperature reading in Ontario for the day (iii) on\text{ontario\textunderscore lowtemp} – lowest temperature reading in Ontario for the day; (iv) on\text{ontario\textunderscore highhumid} – highest humidity reading in Ontario for the day; (v) on\text{ontario\textunderscore avghumid} – average humidity reading in Ontario for the day; (vi) on\text{ontario\textunderscore lowhumid} – lowest humidity reading in Ontario for the day; (vii) on\text{ontario\textunderscore 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 a different variable. With the lags completed, the final incident cases dataset had a total of 174 observations and 60 variables. For this dataset, the response variable was set to the weekly number of PEDV incident cases, with the explanatory variables set as the remaining 59 variables.

To generate the PEDV trend variable for the classification models, a copy of the forecast dataset was made, and a 4-week moving sum of incident cases (with week beginning January 5, 2014) calculated with the zoo rollapply 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. A 4-week moving sum equal to ‘0’ incident cases 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 on the impression that the capacity needed to manage a production-limiting disease relies heavily on industry resources (i.e. industry and a limited number of veterinary practitioners with other daily responsibilities). The weekly count of zero is a preferred option and qualitatively different than 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 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.

3.2.2 Modeling approach

In this study, we aimed to predict the number of new weekly PEDV counts and the
general trend in the number of new cases over a 4-week window, and to subsequently evaluate
the accuracy of such predictions. For the former, we used a regression approach, and for the
latter, we used three different classification approaches.

3.2.2.1 Regression

The random forest algorithm (Breiman, 2001) was selected for the forecasting of new
weekly PEDV counts. The algorithm is a non-parametric predictive modeling method, and
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 is `randomForest` (Liaw and Weiner, 2002) and it provides tuning functions `rfcv` (random forest cross validation for variable selection), and `tuneRF` (tune random forest for the optimal node (mtry) parameter). The `tuneRF` function is useful for ascertaining the number of explanatory variables which should be randomly sampled at each node (mtry), while the `rfcv` function provides the prediction mean square error (regression) or prediction error rate (classification) as predictor variables for a random forest model are sequentially reduced (useful for ascertaining the optimal number of variables for predicting ‘y’). In addition, the `varImpPlot` function (variable importance plot) provides a variable importance measure which ranks each explanatory variable per the percent increase in mean square error (regression), or mean decrease in accuracy (classification), when the variable is randomly permuted and other explanatory variables left unchanged.

For the forecasting of new weekly PEDV counts, caret’s `createdatapartition` and `train` functions in R (Kuhn, 2008) were used to split the dataset with 174 observations into 70% training and 30% testing, with 10-fold cross validation repeated thrice on the training set, and model validation performed on the testing dataset. Based on prior knowledge of contributing factors to PEDV transmission, several random forest models were constructed. Specifically, the following random forest models were constructed with `mtry` values obtained from the `tuneRF` function:

- incident cases ~ infectious sows + infectious sows lags 1 - 5
- incident cases ~ incident cases lags 1 - 5 + prevalence + prevalence lags 1 - 5
- incident cases ~ high temperature + high temperature lags 1 - 5
- incident cases ~ average temperature + average temperature lags 1 - 5
• incident cases ~ low temperature + low temperature lags 1 - 5
• incident cases ~ high humidity + high humidity lags 1 - 5
• incident cases ~ average humidity + average humidity lags 1 - 5
• incident cases ~ low humidity + low humidity lags 1 - 5
• incident cases ~ precipitation + precipitation lags 1 - 5
• incident cases ~ all explanatory variables

Variable importance plot for these models were then generated using the `varImpPlot` function.

3.2.2.2 Classification
The random forest, neural nets, and classification tree algorithms were selected for forecasting PEDV trends. The random forest algorithm was previously described in section 3.2.2.1, and as such, the neural nets and classification tree methods will be covered briefly.

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 neural network implementations in R are `nnet` (Venables and Ripley, 2002) and `neuralnet` (Fritsch and Guenther, 2010). The `nnet` implementation only 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). Variable importance plots for neural net models can be obtained from the neuralnettools olden function (Beck, 2015).

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). Variable importance plots for rpart models can be obtained from the caret dotPlot function (Kuhn, 2008).

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

For the forecasting of PEDV trends, the trend variable was defined as a factor in R (with levels “high”, “medium”, “low”, and “zero”), to ensure the trend variable would be interpreted as a categorical variable. Using caret’s createdatapartition and train functions, the dataset with 171 observations was subsequently split into 70% training and 30% testing, with 10-fold cross validation repeated thrice on the training set, and model validation performed on the testing
dataset. The caret createdatapartition function takes class distribution into account for dataset splits, and as such, ensured a balance of “high”, “medium”, “low”, and “zero” trend observations for the training and test set. All 59 explanatory variables were included in this random forest model (with the mtry value from the tuneRF function), and a variable importance plot was generated. Going forward, this model will be known as the “full” model.

The rfcv function was then applied to the entire dataset, and the results indicated the lowest prediction error rate at 30 variables. As such, all subsequent models were constructed with the top 30 variables from this full model (as obtained from the variable importance plot), with caret’s createdatapartition utilized for the 70% training and 30% testing split, the train function utilized for cross-validation, and the confusionmatrix function utilized for the construction of all confusion matrices, which included model sensitivity and specificity values.

The random forest, neural nets, and classification tree models with the top 30 variables from the full model made use of the dataset with 171 observations. The dataset was split into 70% training and 30% testing, with 10-fold cross validation repeated thrice on the training set, and model validation performed on the testing dataset. The random forest model made use of tuning parameter mtry (from the tuneRF function), and a variable importance plot was generated. For the neural nets model, the dataset was centered and scaled, with nnet size, decay, and maxit parameters set to maximize the classification accuracy on the testing dataset. For classification trees, the minsplit, cp, and xval parameters were set to maximize the classification accuracy on the testing dataset.

To further validate the results obtained with the approaches outlined above, 10-fold cross validation was performed once on the entire dataset. Using caret’s createfolds and train functions, random training and test sets were created for each of the 10 folds, and for each fold,
the training and test sets were fed to a random forest, neural nets, and classification tree model, with explanatory variables and tuning parameters set to the values used previously (i.e. for the corresponding models with a 70-30 split of the data). Unlike the caret createdatapartition function, the caret createfolds function does not take class distribution into account for each fold, and as such, the randomly selected training and sets did not have balanced “high”, “medium”, “low”, and “zero” trend observations.

3.2.2.3 Model Evaluation

For the regression approach, root mean square error (RMSE) values were calculated for each model, with the aim of selecting the best predictive model for new weekly PEDV counts based on the lowest RMSE for actual versus observed values on the test set. For models with RMSE values near to each other, results from the rfcv function were used to select the best performing model. Specifically, the model with explanatory variables closest to the rfcv predictor count with the corresponding lowest mean square error was selected as the best performing model.

For the classification approach with 70% training and 30% test sets, model performance was assessed via confusion matrices, overall classification accuracy, and model sensitivity and specificity values by trend (i.e. “high”, “medium”, “low”, and “zero”), whereas for the classification approach with random training and test sets (10-fold cross validation using the entire dataset), 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.

3.3 Results

For the regression forecasting of new weekly PEDV counts, the low temperature, average
temperature, and the model with all co-variates had the lowest RMSE values (at 1.640, 1.641, and 1.684, respectively). However, since the rf.cv function indicated the optimal number of predictors at 59 explanatory variables, the model with all co-variates was selected as the best performing model.

For the classification forecasting of PEDV trends, the confusion matrix in Table 3.1 provides accuracy measures on the static 70-30 split of the dataset, with random forest, 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 3.2 indicates overall accuracy values of 68%, 57%, and 55% for random forest, neural nets, and classification trees respectively. 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 is presented in Table 3.3, while the boxplot of sensitivity and specificity values across 10 folds on the entire dataset is presented in Figure 3.1. The boxplot of sensitivity and specificity values indicate higher median values for the
random forest model across all 10-folds, the only exception being trends “zero” and “low” for sensitivity and specificity respectively, where neural nets had a higher median value.

Variable importance plots for the random forest regression, random forest classification, neural nets, and classification tree models are presented in Figures 3.2, 3.3, 3.4 and 3.5 respectively. The scatterplot of actual vs predicted PEDV counts for random forest regression is shown in S-Figure 1, while alternate views of variable importance for random forest regression and classification (as rendered by the caret VarImp function) are provided in S-Figures 3.2 and 3.3. A description of variable names for the 30 explanatory variables in all classification models is provided in Table 4.

When considering short-term prediction (i.e. PEDV weekly counts), the top 5 variables from the random forest regression plot (Figure 3.2) indicate that prevalence, number of infectious sows, and average temperature are important. The prevalence 5 weeks prior (lag5_prevalence) is ranked higher than the current week's prevalence. Similarly, the number of infectious sows 4 and 5 weeks prior (lag4_infectious_sows and lag5_infectious_sows) are ranked higher than the current weekly count of infectious sows.

For long-term prediction (i.e. PEDV trend classification), each model ranked the importance of explanatory variables differently. For random forest classification, prevalence was the driving force (Figure 3.3), however, ont_mean_lowtemp is also highly ranked. It appears the current week's low temperature, alongside prior and current prevalence values, determines the PEDV trend 4 weeks into the future. In terms of the actual trend, caret's variable importance plot in S-Figure 3.3 ranks prevalence as the most important predictor for trend "zero", and ont_meanlowtemp as the most important predictor for trend "high".

For neural nets classification, the variable importance is segregated by trend (Figure 3.4).
For both trend “zero” and “high”, the number of infectious sows and prevalence, as well as ont_meanlowtemp, ont_meanavgtemp, and ont_meanhightemp are highly ranked. For trend “zero”, the plot shows a positive relationship for the temperature variables and a negative relationship for the number of infectious sows and prevalence, while for trend “high”, the plot shows a negative relationship for the temperature variables and a positive relationship for the number of infectious sows and prevalence. In other words, as the temperature increases, there is an increased likelihood that the PEDV trend over the next 4 weeks will be "zero", and as the prevalence and number of infectious sows increase, there is a decreased likelihood that the PEDV trend will be “zero”. The converse applies for trend “high”, where a temperature increase results in a decreased likelihood for this trend, and an increase in the number of infectious sows and prevalence results in an increased likelihood for this trend.

For classification trees, the variable importance plot (Figure 3.5) indicates that low, average, and high temperatures (ont_meanlowtemp, ont_meanavgtemp, and ont_meanhightemp) are the top predictors for the PEDV trend, with prior and current prevalence also noted as strong indicators.

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 3.4, 3.5, 3.6, 3.7, and 3.8 respectively. Each plot shows a strong seasonality component, however, there are notable differences where the trend is concerned. For example, incident cases (S-Figure 3.4) started with a sharp trend decrease at the beginning of 2014 but then followed with a gentle decrease, while prevalence (S-Figure 3.5) had a gentle trend increase at the beginning of 2014 followed by a sharp decrease. Both low and average temperature (S-Figures 3.6 and 3.7) show a gradual but strong trend increase, while average humidity (S-Figure 3.8)
started with a gentle trend decrease but then switched to a trend increase in mid-2016.

3.4 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 2,754 premises as confirmed positive for PEDV, with a cumulative total of 3,379 PEDV-positive premises since reporting began in June 2014 (USDA, 2017). The University of California Davis also provides a Disease BioPortal dashboard with a map of the PEDV spread across the United States. As of August 22 2017, PEDV-PCR positive cases were reported in virtually all of the lower 48 states (University of California Davis, 2017). Furthermore, the USDA continues to mandate that all cases of PEDV, whether confirmed or presumed, be reported to state animal health officials, state veterinarians, or a USDA veterinary services district office (USDA, 2017).

The PEDV incursion into Canada was less severe, and though 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 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 Canada, the recent outbreaks in Manitoba (spanning May – August 2017) are a grave reminder that PEDV is still present, and underscores the importance of monitoring, surveillance, and intervention strategies in controlling and possibly eliminating the disease. To this end, forecasting methods can be added to the
surveillance toolkit as an early warning system of potential outbreaks.

For the above-noted PEDV forecast models with a static 70-30 split of the dataset, the neural nets model is the best-performing model, 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., 2010). However, judging by the overall accuracy and the misclassification error, the random forest model is a close second, as it reported just one additional misclassification error than 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 - classification trees work best when a horizontal or vertical split of the explanatory variables also result in accurate split of the response categories (Shmueli et al., 2010).

For the random training and test sets (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. With fewer opportunities to “learn” what constitutes a “high” trend (due to the randomized training and test sets), there’s a notable decrease in overall accuracy for both random forest and neural nets, however, the decrease is marginal for the former and pronounced for the latter. In addition, while classification trees reported an increase in overall accuracy, it was still the worst performing model - further confirmation that this predictive method is not well-suited to the
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 classes “zero” and “low”, and as such provided better sensitivity and specificity values.

Overall, the PEDV trend classification models have much higher specificity values than 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, the cost of a false negative (i.e. false “non-zero” trend) far 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 still not outstanding 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 have an initial sustained peak (usually at the start of the epidemic), after which the epidemic is 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 probably inadequate for the training needs of machine learning models (e.g., random forest, neural nets, and classification trees).

In terms of variable importance for short-term PEDV prediction (i.e., weekly PEDV counts), it appears past disease burden (lagged prevalence and lagged count of infectious sows) is just as important, if not more important, than the current disease burden when it comes to weekly PEDV counts, a result which may well be an indicator of the inherent difficulties involved in PEDV elimination. Furthermore, the highly-ranked average temperature at 3 weeks prior (lag3_ont_meanavgtemp) may be an indicator of the PEDV transmission time from an infected farm to a susceptible one, either due to direct contact (i.e., animal movement associated with the swine production system, from sow to nursery to finisher) or indirect contact (i.e., airborne transmission, transport vehicles, and contaminated fomites).

For long-term PEDV prediction (i.e. PEDV trends), the highly-ranked temperature, lagged prevalence, and current prevalence 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 different 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 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. Thus, the alternative explanation is that veterinary practitioners have become successful in quick elimination of PEDV infection from well-managed sow herds, which could lead to elimination from additional downstream sites. For further insight, studies that address the average time to PEDV elimination are needed. 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. Therefore, the results may not be extrapolated directly to target populations consisting of sow herds only.

Furthermore, it is also important to note that since PEDV is quite stable at low temperatures, PEDV biosecurity measures are more difficult to implement in cold weather. 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), production of disinfectants which maintain their efficacy at low temperatures (Ferry and Benjamin, 2015), and funding for the costs associated with enhanced biosecurity for transport vehicles (OSCIA, 2014).

For the neural nets variable importance measure, the inverse relationship between some lagged variables and trends (e.g. lag5_infectious_sows, lag5_prevalence, and trend “high”) could be attributed to quick PEDV elimination by veterinary practitioners (as explained above), but
may also reflect a key neural network limitation, specifically, that “if the network sees only cases in a certain range, its predictions outside this range can be completely invalid” (Shmueli et. al, 2010). Since the dataset has few observations of trend “high” and many observations of trend “low” and “zero”, it appears this limitation of neural networks applies here.

3.5 Conclusion

This study investigated the use of predictive analytics for the short-term prediction of weekly PEDV incident cases and long-term prediction of PEDV trends. The results show that the random forest regression model with all 59 explanatory variables is the best model for forecasting weekly PEDV incident cases, while the random forest classification model with 30 explanatory variables is the best model for forecasting future PEDV trends. Furthermore, the variable importance measures for random forest models confirm prevalence and temperature as important contributors to future PEDV cases and trends.
References


USDA - United States Department of Agriculture Animal and Plant Health Inspection Service, 2017: Swine Enteric Coronavirus Diseases (SECD), including Porcine Epidemic Diarrhea


**Tables**

Table 3.1: PEDV trend classification performance for Ontario PEDV and weather data from January 2014 - April 2017 (171 observations), with 70% of the dataset allocated for model training and 30% allocated for model testing

<table>
<thead>
<tr>
<th>Reference</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Zero</th>
<th>Overall Accuracy</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance - random forest model with 30 co-variates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Zero</td>
<td>Overall Accuracy</td>
<td>95% CI</td>
</tr>
<tr>
<td>High</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.71</td>
<td>0.57 – 0.83</td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Performance - neural nets model with 30 co-variates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Zero</td>
<td>Overall Accuracy</td>
<td>95% CI</td>
</tr>
<tr>
<td>High</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.76</td>
<td>0.61 – 0.87</td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Performance - classification tree model with 30 co-variates</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Zero</td>
<td>Overall Accuracy</td>
<td>95% CI</td>
</tr>
<tr>
<td>High</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0.45</td>
<td>0.31 – 0.60</td>
</tr>
<tr>
<td>Medium</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2: 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) for model training and testing

<table>
<thead>
<tr>
<th>Reference</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Zero</th>
<th>Prediction</th>
<th>Overall Accuracy</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-fold cross validation performance - random forest model with 30 co-variates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Overall Accuracy</td>
<td>95% CI</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Zero</td>
<td>Prediction</td>
<td>Overall Accuracy</td>
<td>95% CI</td>
</tr>
<tr>
<td>High</td>
<td>11</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0.68</td>
<td>0.60 – 0.75</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>2</td>
<td>33</td>
<td>5</td>
<td>2</td>
<td>0.57</td>
<td>0.49 – 0.64</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1</td>
<td>9</td>
<td>34</td>
<td>11</td>
<td>0.55</td>
<td>0.47 – 0.63</td>
<td></td>
</tr>
<tr>
<td>Zero</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>38</td>
<td>0.55</td>
<td>0.47 – 0.63</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.3: PEDV trend classification diagnostics for Ontario PEDV and weather data from January 2014 - April 2017, with 70% training and 30% test sets, as well as randomly allocated training and test sets (10-fold cross validation) for model training and testing.

<table>
<thead>
<tr>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.60</td>
<td>0.75</td>
<td>0.69</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>0.95</td>
<td>0.82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.80</td>
<td>0.58</td>
<td>0.75</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95</td>
<td>1.00</td>
<td>0.85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.40</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.93</td>
<td>0.81</td>
<td>0.67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.58</td>
<td>0.79</td>
<td>0.62</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.98</td>
<td>0.87</td>
<td>0.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.47</td>
<td>0.60</td>
<td>0.53</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.96</td>
<td>0.79</td>
<td>0.75</td>
</tr>
</tbody>
</table>
Table 3.4: A description of the 30 explanatory variables used in all PEDV trend classification models (PEDV long-term prediction), as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017

<table>
<thead>
<tr>
<th>variable name</th>
<th>timeframe</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>infectious_sows</td>
<td>current week</td>
<td>number of PEDV-infectious sows at the end of the week</td>
</tr>
<tr>
<td>ont_meanavgtemp</td>
<td></td>
<td>mean of daily average temperature readings</td>
</tr>
<tr>
<td>ont_meanhightemp</td>
<td></td>
<td>mean of daily high temperature readings</td>
</tr>
<tr>
<td>ont_meanolowtemp</td>
<td></td>
<td>mean of daily low temperature readings</td>
</tr>
<tr>
<td>prevalence</td>
<td></td>
<td>PEDV prevalence at the end of the week</td>
</tr>
<tr>
<td>lag1_infectious_sows</td>
<td>1 week prior to the current week</td>
<td>number of PEDV-infectious sows at the end of the week</td>
</tr>
<tr>
<td>lag1_ont_meanavghumid</td>
<td></td>
<td>mean of daily average humidity readings</td>
</tr>
<tr>
<td>lag1_ont_meanavgtemp</td>
<td></td>
<td>mean of daily average temperature readings</td>
</tr>
<tr>
<td>lag1_ont_meanhightemp</td>
<td></td>
<td>mean of daily high temperature readings</td>
</tr>
<tr>
<td>lag1_ont_meanolowtemp</td>
<td></td>
<td>mean of daily low temperature readings</td>
</tr>
<tr>
<td>lag1_prevalence</td>
<td></td>
<td>PEDV prevalence at the end of the week</td>
</tr>
<tr>
<td>lag2_infectious_sows</td>
<td>2 weeks prior to the current week</td>
<td>number of PEDV-infectious sows at the end of the week</td>
</tr>
<tr>
<td>lag2_ont_meanavgtemp</td>
<td></td>
<td>mean of daily average temperature readings</td>
</tr>
<tr>
<td>lag2_ont_meanhightemp</td>
<td></td>
<td>mean of daily high temperature readings</td>
</tr>
<tr>
<td>lag2_ont_meanolowtemp</td>
<td></td>
<td>mean of daily low temperature readings</td>
</tr>
<tr>
<td>lag2_prevalence</td>
<td></td>
<td>PEDV prevalence at the end of the week</td>
</tr>
<tr>
<td>lag3_infectious_sows</td>
<td>3 weeks prior to the current week</td>
<td>number of infectious sows at the end of the week</td>
</tr>
<tr>
<td>lag3_ont_meanavgtemp</td>
<td></td>
<td>mean of daily average temperature readings</td>
</tr>
<tr>
<td>lag3_prevalence</td>
<td></td>
<td>prevalence at the end of the week</td>
</tr>
<tr>
<td>lag4_infectious_sows</td>
<td>4 weeks prior to the current week</td>
<td>number of PEDV-infectious sows at the end of the week</td>
</tr>
<tr>
<td>lag4_ont_meanavgtemp</td>
<td></td>
<td>mean of daily average temperature readings</td>
</tr>
<tr>
<td>lag4_ont_meanhightemp</td>
<td></td>
<td>mean of daily high temperature readings</td>
</tr>
<tr>
<td>lag4_ont_meanolowtemp</td>
<td></td>
<td>mean of daily low temperature readings</td>
</tr>
<tr>
<td>lag4_prevalence</td>
<td></td>
<td>PEDV prevalence at the end of the week</td>
</tr>
<tr>
<td>lag5_infectious_sows</td>
<td>5 weeks prior to the current week</td>
<td>number of PEDV-infectious sows at the end of the week</td>
</tr>
<tr>
<td>lag5_ont_meanavgtemp</td>
<td></td>
<td>mean of daily average temperature readings</td>
</tr>
<tr>
<td>lag5_ont_meanhightemp</td>
<td></td>
<td>mean of daily high temperature readings</td>
</tr>
<tr>
<td>lag5_ont_meanolowtemp</td>
<td></td>
<td>mean of daily low temperature readings</td>
</tr>
<tr>
<td>lag5_prevalence</td>
<td></td>
<td>PEDV prevalence at the end of the week</td>
</tr>
</tbody>
</table>
Figures

Figure 3.1: 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
Figure 3.2: Variable importance plot for the random forest regression model with all co-variates (PEDV short-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017
Figure 3.3: 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.
Figure 3.4: Variable importance plot for the neural nets classification model with 30 co-variates (PEDV long-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017.
Figure 3.5: Variable importance plot for the classification tree model with 30 co-variates (PEDV long-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017
Supplementary Material

S-Figure 3.1: Scatterplot of actual vs predicted weekly PEDV counts for the random forest regression model with all co-variates (PEDV short-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017
S-Figure 3.2: An alternate view of variable importance (rendered by the caret VarImp function) for the random forest regression model with all covariates (PEDV short-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017
S-Figure 3.3: An alternate view of variable importance (rendered by the caret VarImp function) for the random forest classification model with 30 covariates (PEDV long-term prediction), as applied to the Ontario PEDV and weather dataset for January 2014 - April 2017
S-Figure 3.4: Time series decomposition plot for the weekly count of PEDV incident cases, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017
Figure 3.5: Time series decomposition plot for the weekly PEDV prevalence, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017
Figure 3.6: Time series decomposition plot for the weekly low temperature, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017
S-Figure 3.7: Time series decomposition plot for the weekly average temperature, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017
S-Figure 3.8: Time series decomposition plot for the weekly average humidity, as obtained from the Ontario PEDV and weather dataset for January 2014 - April 2017
CHAPTER FOUR – Research Summary and Conclusions

As swine production systems evolve to meet the increasing global demand for pork products, novel swine pathogens are expected to emerge (Fournie et al., 2012). The zoonotic potential for such pathogens and the emergence of virulent strains of existing pathogens have global trade implications and could result in severe economic losses for producers. As a result, active disease monitoring and surveillance will continue to play a critical role in detecting abnormal disease frequencies for known pathogens, the emergence of novel pathogen, or the emergence of known pathogens in regions previously free of the disease, as was the case with Porcine Epidemic Diarrhea Virus (PEDV) in North America.

Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) are of economic importance to producers, however, few studies have been published about the disease burden of both pathogens (Wang et al., 2016; Zhang, 2016; Romana et al., 2016; Chae et al., 2000; Carvajal et al., 1995). Furthermore, while health and disease forecasting with machine learning methods have gained considerable traction in human medicine, traditional statistical methods are still more prevalent in veterinary medicine, although some studies, notably in diseased chicken identification (Yang et al., 2005), infection and disease in California sea lions (Mancia et al., 2012), fish disease diagnosis (Zeldis and Prescott, 2010), disease susceptibility in cows (Naderi et al., 2016), diagnosis of avian diseases (Banakar et al., 2016), and livestock disease surveillance (Yazdanbakhsh et al., 2017), have utilized machine learning methods.

This thesis investigated the incidence and prevalence of Porcine Epidemic Diarrhea Virus (PEDV) and Porcine Deltacoronavirus (PDCoV) in Ontario between 2014 - 2016, thereby contributing to the body of knowledge regarding the disease burden of these pathogens in Canada. The analysis of PEDV and PDCoV surveillance data in an industry database revealed
declining incidence risk, incidence rates, and prevalence for both pathogens in Ontario. Specifically, for 2014, 2015, and 2016, the estimated PEDV incidence risk was 13.49%, 2.97%, and 1.42%, incidence rate was 0.14, 0.03, and 0.02 (cases per herd year), and prevalence was 4.36%, 2.25%, and 1.35%. For PDCoV, the estimated incidence risk for 2014, 2015, and 2016 was 1.14%, 0.30%, 0.08%, incidence rate was 0.011, 0.003, and 0.001 (cases per herd year), and prevalence was 0.48%, 0.17%, and 0.16%. Visualizations for these measures were provided on a monthly basis for each year, and the research also highlighted similarities and differences in PEDV reporting provided by the province (Ontario Ministry of Agriculture Food and Rural Affairs) and the industry (Ontario Swine Health Advisory Board).

The research went a step further by exploring the use of machine learning methods – specifically Random Forest, Artificial Neural Networks, and Classification and Regression Trees (CART) - for PEDV disease forecasting. PEDV incidence and prevalence, as well as weather data (temperature, humidity, and precipitation), were supplied to each model to determine the best method for long-term forecasting of PEDV trends as well as the impact of weather variables on PEDV. The study found that the Random Forest model provided the best prediction for future PEDV trends, both in terms of classification accuracy and sensitivity/specificity values, and also revealed that low temperatures were strongly associated with PEDV incidence. A similar study utilizing incidence data for the prediction of avian influenza H5N1 outbreaks found that the Random Forest model outperformed the ARIMA model in terms of predictive ability (Kane et al., 2014).

There are certain limitations to this research. First, there is the danger of under-reporting for the disease frequency measures, since surveillance data in the industry database was collected within the framework of a voluntary disease control program (DCP). In addition, PEDV and
PDCoV are no longer considered reportable diseases in Ontario, so it is likely that some cases were not captured in the provincial or industry-administered surveillance systems. Furthermore, since presumed PEDV and PDCoV positive herds were not tested for the virus, some misclassification is likely as some herds may be negative for PEDV or PDCoV but positive for pathogens with similar clinical signs, such as transmissible gastroenteritis virus (TGEV).

For forecasting, most machine learning methods require a large dataset spanning long periods (preferably thousands of observations collected over several years) in order to effectively learn patterns in the data and also for generating accurate forecasts. In this instance, the emerging nature of PEDV meant there were few data points to work with (roughly 170+ observations) and the dataset was highly imbalanced with respect to the dependent variable (i.e. PEDV incident cases or trends). In addition, the thesis did not investigate auto-correlations and cross-correlations which are inherently present since each variable represents a time series. Such investigation may have reduced the number of variables, perhaps resulting in more parsimonious models and better prediction accuracy. Further research into the use of data-driven methods for small datasets, such as the study conducted by Pasini (2015) for artificial neural networks, are needed. While techniques such as over-sampling and under-sampling have been proposed for dealing with imbalanced datasets (Altini, 2015), peer-reviewed studies are needed to validate these techniques for use with incidence and prevalence data.

Despite the above-noted limitations, this thesis confirms that surveillance data, when properly analyzed, provides insights into both current and future outbreak scenarios for a disease. For swine diseases, such scenarios are especially important given the expected rise in novel pathogens. Similar studies can be conducted on swine surveillance data maintained by other industry groups, such as Saskatchewan Pork, Manitoba Pork, and Pork Development Center of
Quebec (CDPQ), in order to develop a national picture of PEDV and PDCoV in Canada.
References


