

**University of Alberta**

Comparing outlier detection methods to improve the quality of citizen science data

by

Jennifer Shil-Mun Li

A thesis submitted in partial fulfillment of the requirements for a dual degree

Master of Forestry, University of Alberta, Edmonton, Alberta

and

Master of Science in Agriculture and Forestry, University of Eastern Finland

© Jennifer Shil-Mun Li

Spring 2018

Edmonton, Alberta

The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

## ABSTRACT

Citizen science is the public participation in research, usually through volunteer monitoring or data collection. Data collected by citizen scientists is a valuable resource in many fields of research that require long-term observations across broad spatial scales. However, such data may not be as accurate as those collected by trained professionals. The objective of this thesis is to analyze the reliability of individual observers and observations to enhance the data quality of a citizen science network that has recorded plant phenology (bloom times) since 1987 across Alberta to track biological response to environmental change. This study evaluates five algorithms designed to detect outlier observations and inconsistent observers. These methods rely on different quantitative approaches, including residuals of linear models, correlations among observers, and deviations from multivariate clusters, and percentile-based outlier removal. The effects of these data pre-processing approaches was evaluated by comparing regional means of the resulting time series, through comparing maps of observations that were removed by different methods, and by evaluating spatial autocorrelations, measured as Moran's I. Spatial autocorrelations are expected to increase if outliers and inconsistent observations are successfully removed. All data cleaning methods resulted in an improvement of Moran's I statistic, with percentile-based outlier removal and the clustering method showing the greatest increase in autocorrelations. Methods based on residual analysis of linear models had the strongest impact on the final bloom time mean estimates, but were among the weakest based on autocorrelation analysis. Removing entire sets of observations from potentially unreliable observers proved least effective. In conclusion, percentile-based outlier removal emerges as a simple and effective method to improve reliability of citizen science phenology observations.

## ACKNOWLEDGEMENTS

This thesis was submitted in partial fulfillment of the requirements for a dual-degree for Master of Forestry, University of Alberta, Edmonton, Alberta, and Master of Forest Sciences, University of Eastern Finland.

The process developing this thesis has been one of the most challenging yet rewarding processes that I have ever embarked on, and would not have been possible without the guidance of my supervisors Dr. Andreas Hamann at the University of Alberta, and Dr. Heli Peltola at the University of Eastern Finland. Thank you for the time you have dedicated to answering my questions, your advice in developing this project, and also for challenging me with new ideas and different ways to view my research question. Your knowledge and expertise have guided me throughout this research and report writing process.

A special thank you goes out to Dr. Elisabeth Beaubien, PlantWatch Alberta coordinator, at the University of Alberta. Without your ongoing dedication to the PlantWatch program, this project would not have been possible. I appreciate your inspiration in finding new ideas to pursue, and for your support in the thesis-writing process. Your suggestions and recommendations have been invaluable in developing my thesis drafts.

To my lab-mates at the University of Alberta, thank you for always being willing to help with any scripting problems, and for the company and comradery that made the research environment fun and encouraging.

I extend my gratitude to all the citizen scientists involved in PlantWatch. This project would not have been possible without your time and energy over the years.

To my family, my soon-to-be family, and all of my friends, thank you for your willingness to listen, and also for your encouragement throughout my continued schooling.

The Trans-Atlantic Forestry Master (TRANSFOR-M) program has been an amazing experience. I have had the chance to meet and befriend exceptional people, and had the opportunity to learn from inspiring individuals on both sides of the Atlantic Ocean. Thank you to the program coordinators for the opportunity to partake in such a one-of-a-kind journey.

## CONTENTS

<b>1. INTRODUCTION</b>	<b>1</b>
1.1 Citizen science and its benefits	1
1.2 Problems and approaches for quality assurance	3
1.3 Research objectives	6
<b>2. MATERIALS AND METHODS</b>	<b>6</b>
2.1 Study layout	6
2.2 Alberta PlantWatch phenological data	8
2.3 Climatic data	10
2.4 Data cleaning methods used for phenological data quality improvement	11
Method 1 - Standardized difference (SDiff)	11
Method 2 - Linear model (LM)	12
Method 3 - Linear model by natural subregion (LMNSR)	13
Method 4 - Observer correlation (CORR)	14
Method 5 Dimensionality reduction and clustering (DRC)	16
2.5 Assessment of validity for different data cleaning methods	18
<b>3. RESULTS</b>	<b>20</b>
3.1 Improvement of overall data quality	20
3.2 Global outlier treatment	21
3.3 Effects of data cleaning on regional means	22
3.4 Spatial patterns for removal of records due to data cleaning	23
3.5 Similarity of outlier rankings between data cleaning methods	25
3.6 Other findings for specific methods	26
<b>4. DISCUSSION</b>	<b>29</b>
<b>5. CONCLUSIONS</b>	<b>31</b>
<b>6. LITERATURE CITED</b>	<b>33</b>
Appendix A – Data trends before and after cleaning by species and phase	38
Appendix B – Moran’s I for each cleaning method by species and phase	40

## LIST OF TABLES

<b>Table 1.</b> Conceptual examples of spatial autocorrelation .....	19
<b>Table 2.</b> Two-way analysis of variance (ANOVA) results of Moran’s I statistic values, with individual year (1987-2016) and species phase (4 species x 2 phases) as random effects. Method factors include the original dataset as well as five assessed methods of data cleaning (5% of records were removed during data cleaning). .....	20
<b>Table 3.</b> Change in predicted date for phenological occurrence date for natural subregions after data cleaning compared to the original predicted date for phenological occurrence. ....	22
<b>Table 4.</b> Illustration of correlations, and statistical strengths of correlations between removal values amongst five assessed data cleaning methods Method 1 Standardized Difference (SDiff), Method 2 Linear model (LM), Method 3 Linear model by natural subregion (LMNSR), Method 4 Observer correlation (CORR), and Method 5 Dimensionality reduction and clustering (DRC). .....	26
<b>Table 5.</b> Proportion of removed records through Method 4 (Observer correlation) data cleaning. A mixed approach to data cleaning was used based on the number of records submitted by any individual observer within any given natural subregion. In total, 5% of the overall records were removed for each species phase. ....	27
<b>Table 6.</b> Fitted dimensionality reduction and clustering models from Method 5: Dimensionality reduction and clustering technique for automated data cleaning. ....	28

## LIST OF FIGURES

<b>Figure 1.</b> Environmental context for the province of Alberta, Canada.....	7
<b>Figure 2.</b> Natural subregions for ecological classifications used for data cleaning Method 1 Standardized difference, Method 3 Linear model by natural subregion, and Method 4 Observer correlation. ....	8
<b>Figure 3.</b> Observations made by Citizen Scientists for the Alberta PlantWatch program from 1987 to 2016. Observation counts and natural subregions are depicted with matching colours.	9
<b>Figure 4.</b> Observations made for the Alberta PlantWatch program from 1987 to 2016 used as a part of the data cleaning study. Observer counts and locations are depicted with matching colours.....	10
<b>Figure 5.</b> Conceptual depiction of Method 1 (Standardized difference) data cleaning based on the removal of outliers grouped by natural subregion and year. Blue represents conceptually retained records, red represents conceptually removed records. ....	12
<b>Figure 6.</b> Conceptual depiction of Method 2 (Linear model) data cleaning, based on the development of a linear model with latitude, longitude, elevation, and year as predictors. Observations with the largest residual difference are removed. Blue represents conceptually retained records, red ‘x’ represents conceptually removed records.....	13
<b>Figure 7.</b> Conceptual depiction of Method 3 (Linear model by natural subregion) data cleaning, based on the development of a linear model with natural subregion and year as predictors. Observations with the largest residual difference are removed. Blue represents conceptually retained records, red ‘x’ represents conceptually removed records.....	14
<b>Figure 8.</b> Conceptual depiction of Method 4 (Observer correlation) data cleaning, based on the removal of observers with the lowest correlation to the average observations by natural subregion. Blue dashed lines represents conceptually retained observers; red solid line represents a conceptually removed observer. ....	15

**Figure 9.** Conceptual depiction of Method 5 (Dimensionality reduction and clustering) data removal method. T-distributed stochastic neighbor embedding dimensionality reduction reduces data to two dimensions and clustered into groups (left panel). The points with the largest standardized difference to its cluster mean are removed. Blue represents conceptually retained records, and red ‘x’ represents conceptually removed records (right panel). .....17

**Figure 10.** Effectiveness of data cleaning methods as measured by Moran’s I statistic ( $\pm$ SE) before and after data cleaning. N=240 groups of datasets per cleaning method (4 species x 2 phases x 30 years). 5% of records were removed during data cleaning for each data cleaning method. The “full dataset” represents the Moran’s I statistic for the original records prior to data cleaning .....20

**Figure 11.** Differences among data cleaning methods in the treatment of global outliers for five assessed data cleaning methods. Outliers were classified as having values beyond the interquartile range (the 25th percentile and 75th percentile) multiplied by 1.5.....21

**Figure 12.** Spatial locations of removed records for assessed data cleaning methods. For Methods 1 and 5, the size of the circle represents the absolute difference from the mean. For Methods 2 and 3, the size of the circle represents the absolute difference from the model, and for Method 4, the size of the circle represents the correlation of removed observers multiplied by -1. ....24

**Figure 13.** Automated clusters generated through Method 5 (Dimensionality reduction and clustering). Climatic variables, location, and observed phenological occurrence date were incorporated for automated dimensionality reduction and clustering. Each colour represents and individual cluster grouping. ....28

# 1. INTRODUCTION

## 1.1 Citizen science and its benefits

Citizen science is broadly defined as “scientific study that includes non-professional scientists as contributors or collaborators” which may include sample collection with standardized protocols. Citizen science has been documented as early as 3 500 years ago with citizens and officials recording locust outbreaks in China (Miller-Rushing et al 2012). In Europe, citizen science networks became an important resource for scientific progress in the 17<sup>th</sup> century for natural history observations. For example, early ecologists such as John Ray and Carl Linnaeus have relied extensively on observations reported and specimens collected by amateur naturalists (Miller-Rushing et al 2012).

Today, volunteer observers contribute to various research fields, including astronomy, conservation science, population ecology, environmental risk assessments, pollution detection, and monitoring environmental change (Fuccillo et al 2014; Mengersen et al 2017). Their contributions enable large-scale scientific data collection efforts that would otherwise not be possible. For example, national-scale programs, such as the French Breeding Bird Survey, running since 1989, have been estimated to cost hundreds of thousands to several million Euros per year if the same work were to be carried out by paid professionals instead of volunteers (Levrel et al 2010). Similarly, volunteer contributions to the Cornell University’s FeederWatch program, located in the United States, have been valued at 3 million dollars per year (Dickinson et al 2010). Even for smaller scientific studies, citizen science contributions are often invaluable. For example, in a taxonomic study of ladybird beetles (*Coccinellidae spp.*), volunteer contributions were estimated to be worth several tens of thousands of dollars (Gardiner et al 2012).

In general, any type of biological or environmental monitoring over large geographic areas or long time periods tends to benefit from citizen science networks. For example, citizen science driven projects have been used to identify pollution sources (McKinley et al 2017). The establishment, spread, and control of invasive species are regularly supported by volunteer observation networks (Crall et al 2015). Citizen scientists monitor dozens of invasive plants in Portugal (Marchante et al 2017). In Italy, invasive species and their effects on native species are



monitored by volunteers (Buldrini et al 2015). Also, invasive animals are reported and monitored through citizen science networks (Anderson et al 2017; Morii and Nakano 2017). In conservation biology, rare plant populations are monitored, and potential threats to populations have been identified through data collected by volunteers (Havens et al 2012; Vander Stelt et al 2017). The previously mentioned French Breeding Bird Survey has demonstrated a northward-shift of bird communities, as well as changes in bird community composition, coinciding with climate change (Jiguet et al 2012). The results of an annual Christmas bird count in North America, has also found poleward shifts in the range boundaries of birds from 1975 to 2004, potentially due to anthropogenic and climatic factors (La Sorte and Thompson 2007).

In the context of environmental monitoring, perhaps the most important citizen science contribution is found in the field of phenology and climate change. Phenology, is based on “the seasonal timing of life cycle events” (Rathcke and Lacey 1985). Citizen scientist supported plant phenology programs include the USA National Phenology Network (USANPN), which monitors the timing of flowering and leafing of approximately 878 plant species (USA National Phenology Network n.d.). In Canada, the Alberta PlantWatch program is the longest currently running citizen-science plant phenology observation network (Beaubien and Hamann 2011b). In Europe, phenological monitoring programs include the International Phenological Gardens, founded in 1957 to monitor genetically identical plants across the continent, although nationwide phenological monitoring programs are common, for example the former USSR has recorded phenological occurrences since the 1850s, and since the 1920s in Estonia, and portions of Slovak and the Czech Republic (Menzel 2003). In recent decades, data from such phenology monitoring networks have emerged from relative obscurity to the forefront of environmental monitoring. For example, an analysis of published studies in plant phenology within the International Journal of Biometeorology alone have increased from approximately 350 papers per decade between 1957-2007 to over 1 000 contributions between 2007 to 2016 (Donnelly and Yu 2017).

Citizen science networks not only benefit the scientific community, but the benefits can return to the volunteers through their involvement in managing local resources and environments, or through education and promotion of conservation programs. By incorporating the local involvement of citizen scientists, there is potential for citizen science driven projects to support environmental management and policy-making (McKinley et al 2017). In an example of a citizen

scientist-driven research project, the Shermans Creek Conservation Association in Pennsylvania, USA, designed a study with the assistance of researchers from the Alliance for Aquatic Resource Monitoring, and collected and analyzed three years of data relating to the condition of Shermans Creek with the goal of future participation in decision making (Shirk et al 2012). In another example, data collected by recreational fishers was instrumental to determining the boundaries of a Great Barrier Reef Marine Park (Granek et al 2008).

## **1.2 Problems and approaches for quality assurance**

Several criticisms have been raised regarding the reliability and objectivity of citizen scientists' data (Danielsen et al 2014). As unpaid researchers, the motivation and objectivity of citizen scientists have been questioned, and concerns have arisen from policies and decisions derived from citizen scientist-produced data in which the observers have a vested interest in the outcomes of the study (Kosmala et al 2016). In the USA, arguments against the reliability of volunteer data have led to some programs reverting to the use of professional scientists, or to limiting volunteer involvement (Silvertown et al 2013; MacKenzie et al 2017). In 1993, while establishing the United States National Biological Survey (now a part of the Biological Resources Division of the United States Geological Survey), an amendment was made to the original bill prohibiting the use of volunteers in survey activities (Reichhardt 1994; Wagner 1999; Lewis 2003).

Research has provided some evidence that supports the above criticisms regarding quality issues with data obtained from volunteer networks. For observations where temporal accuracy needs to be high, a potential sampling bias includes a “weekender effect”, where reported observations are more frequent on the weekends, potentially causing a delay in reported event dates. This is apparent in particular with migration studies of birds, where an analysis of citizen scientist bird migration data from 1947 to 2004 found that 44% of first arrival reports were made on weekends, as opposed to the expected 28%. Rare birds were also primarily observed on weekends (Sparks et al 2008). Another important sampling bias is based on the geographic distribution of sampling sites, where easily accessible areas are more frequently sampled, and remote areas are less frequently sampled (Hugo and Altwegg 2017).

Errors in citizen science data may also result from incorrect data entry. Records are often recorded manually by volunteers, and are later transcribed into a computer database, where there is possible error due to interpretation of handwriting, and in the manual entry process of data into the database. Incorrect identification of species is also possible and accuracy of observations has been found to decline with increasing difficulty to distinguish among closely related or visually similar species (Beaubien and Hamann 2011b; Crall et al 2011; Fuccillo et al 2014; Kosmala et al 2016). It has been noted earlier that the quality of observations can be influenced in part by the education level of observers (Dickinson et al 2010). However, some studies have contradicted this information (Danielsen 2014). Citizen scientists who are invested in a project, involved in the study design, and properly trained, even in communities with limited abilities to read and write, can produce data similar to that of trained scientists (Danielsen 2014).

Nevertheless, other research has also shown that when issues such as those listed above are properly taken into consideration, the inherent concerns on accuracy and bias in citizen science data may be accounted for at the stage of data preparation. In evaluating the quality of citizen scientist data, a variety of factors should be taken into account, including the level of effort required for a task, the level of training invested in observers, and the study design (Kosmala et al 2016). For example, relatively straight forward observation protocols by the United States National Phenology Network has yielded a higher than 90% match when compared to independent observations made by a professional ecologists, while more difficult to observe transition phases were accurate at approximately 70% (Fuccillo et al 2014). Previous studies of citizen science accuracy have found that citizen scientists could assess phenology of easy to identify plant species with approximately 80% accuracy, declining to 65% for species that could be easily confused with others (Crall et al 2011).

For difficult tasks, careful training of volunteers should be taken into consideration, as well as ongoing assessment, validation, and pre-tests in order to minimize error (Kosmala et al 2016). In previous literature from Alberta PlantWatch, it has been found that citizen scientist driven data is high quality if proper training has been given for observers (Beaubien and Hamann 2011b). The effectiveness of volunteer training can also be evaluated. For example, the global eBird observation network uses machine learning methods to improve observer training, and validate submitted data (Kelling et al 2013). Effective training has also been documented as instrumental

in citizen science driven studies related to water quality assessments and invasive plant monitoring (Fore et al 2001; Gallo and Waitt, 2011). Improvement to training and observation protocols in science is not unique to volunteer networks, but similarly applies to data collection by professionals and experts, where repetitive tasks invariably lead to fatigue and errors in data entry and management (Kosmala et al 2016).

Data cleaning, outlier detection, or other data pre-processing is a common scientific approach that does not only apply to volunteer data, but also to any professionally collected data or instrument measurements that are prone to inaccuracies. Such methods have been routinely applied to improve volunteer data. According to the by the Data Observation Network for Earth (DataONE) – a United States National Science Foundation funded agency to encourage data sharing, the quality of data could be evaluated after data input by using a statistical model for the visual detection of outliers, and subsequent investigation (DataONE n.d.). An example of this statistical model could be based on that of Ranjitkar (2013) where a significant correlation was found between flowering date and location (latitude and elevation) for *Rhododendron arboretum* Sm. in Nepal. Spatiotemporal verification of data has also been conducted in the past on citizen scientist-submitted data, utilizing volunteer validation, expert validation, or geographic assessment of quality (Mehdipoor et al 2015).

Another way that biological data from volunteer networks could potentially be evaluated without a direct comparison to professional observers is the internal consistency of observations, based on the general principle expressed by Tobler's first law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). Plant blooming and budburst in the temperate region is highly dependent on temperature, measured through heat sum accumulation (Rathcke and Lacey 1985), which is spatially autocorrelated (Javari 2017). Factors that influence the timing of budburst and flower blooming may also include plant genetics, and environmental factors such as photoperiod, and moisture in tropical areas (Rathcke and Lacey 1985). These environmental factors vary geographically and are also inherently related to location and elevation. A previous study by Schwartz, Hanes, and Liang (2014) found that using Moran's I, clustering of phenological observations occurred across the study area of 625 m x 625 m. At smaller (microclimate) scales within the larger study area, Schwartz et al (2014) found that spatial autocorrelation measures depicted random patterns.

### **1.3 Research objectives**

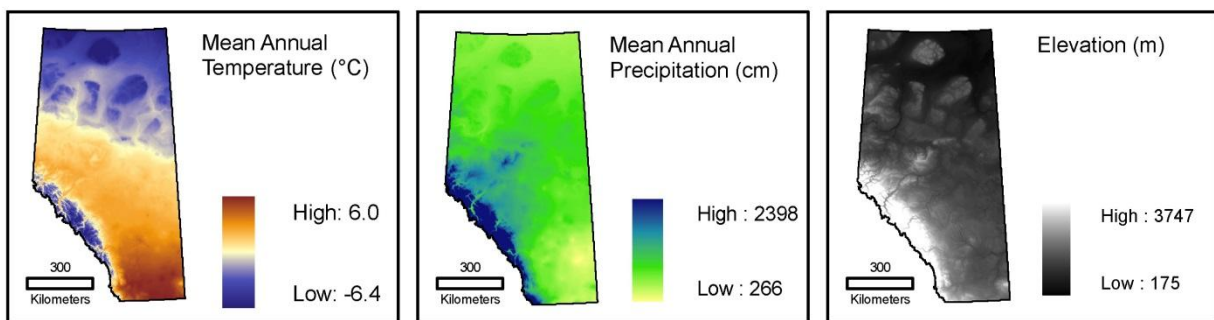
This study evaluates algorithms designed to detect outlier observations and inconsistent observers from a citizen science network that monitors the timing of bloom and leaf out for 30 plant species in Alberta, Canada. The network consists of 700 observers that have reported more than 50,000 bloom dates from 1987 to 2016. The data has been used to document the impact of climate change at northern latitudes (Beaubien and Johnson 1994; Beaubien and Freeland 2000; Beaubien and Hamann 2011a). Data from Beaubien and Freeland 2000 has been featured in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change as evidence for the impacts of climate change (IPCC 2007). Due to the large number of observers, the geographic extent of the network over more than a thousand kilometers, covering a variety of climates and ecosystems, makes outlier detection challenging. This study contributes a comparison of five data cleaning methods that rely on different quantitative approaches, including residuals of linear models, correlations among observers, and deviations from multivariate clusters, and percentile-based outlier removal. As a measure for determining the best method for potentially unreliable observation removal, Moran's I statistic (Moran's I) was used as a measure of spatial autocorrelation to determine if the cleaned datasets had an increase in spatial autocorrelation when compared to the original dataset. Spatial autocorrelations are expected to increase if outliers and inconsistent observations are successfully removed. In addition to the effect of different cleaning methods on spatial autocorrelations, the data cleaning approaches were also evaluated by comparing regional means of the resulting time series, and through comparing maps of observations that were removed by different methods.

## **2. MATERIALS AND METHODS**

### **2.1 Study layout**

This study was implemented in Alberta, one of Canada's three Prairie Provinces, and is bound to the north by the 60<sup>th</sup> parallel, the 110<sup>th</sup> meridian to the east, and the 120<sup>th</sup> meridian and the Rocky Mountains to the west (Smith et al 2017). With an area of 661 848 km<sup>2</sup>, topography ranges from mountainous regions to the west with elevations up to 3 747m, and slopes downwards towards

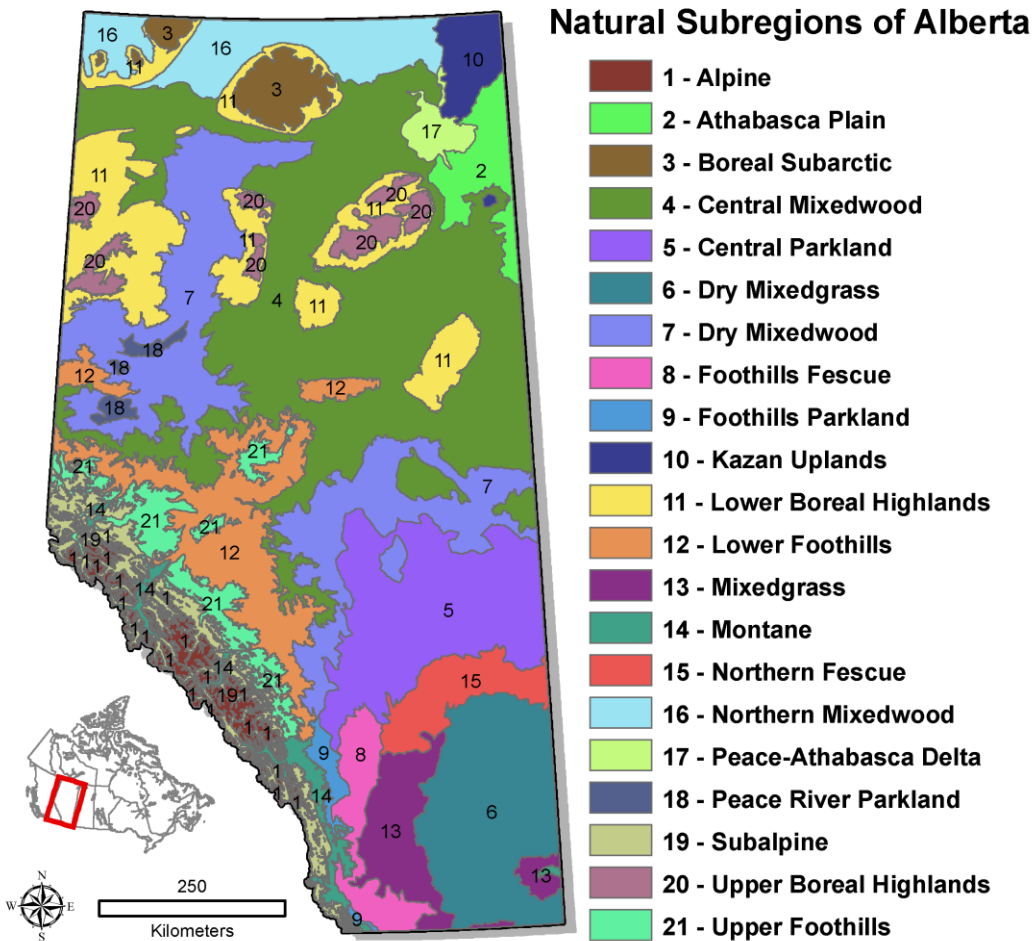
the northeast where the lowest elevation is 175m (Figure 1) (Alberta Environment and Parks 2015, Smith et al 2017). Temperatures also fluctuate annually from summer daytime high temperatures averaging 20 to 25°C, to winter daytime low temperatures averaging -5 to -15°C (Travel Alberta 2017). Climatic regimes in Alberta include the grassland region in the southeast, characterized by a continental climate of long cold winters and low precipitation (Strong and Legatt 1992). The cordilleran climatic region encompassing the Rocky Mountains is located to the southwest of the province and has short summers and variable winters (Strong and Legatt 1992). The boreal region in the north of the province has low precipitation, but long daylight hours in summer (Strong and Legatt 1992). As outlined by the Natural Regions Committee (2006), Alberta is divided into six natural regions based on landscape pattern (vegetation, soils, and physiographic features).



**Figure 1.** Environmental context for the province of Alberta, Canada.

Datasets including the Alberta Climate Model developed by Alberta Environment (2005) and the provincial digital elevation model were incorporated with field surveys to develop the natural regions of Alberta (Figure 2). These natural regions are further subdivided into 21 subregions, and “generally characterized by vegetation, climate, elevation, and latitudinal or physiographic differences” within regions.

Given the scale of Alberta and the wide range of geographic variability, natural subregions were selected for this study as a means of landscape classification in order to generalize areas of Alberta that are expected to have similar phenological responses.

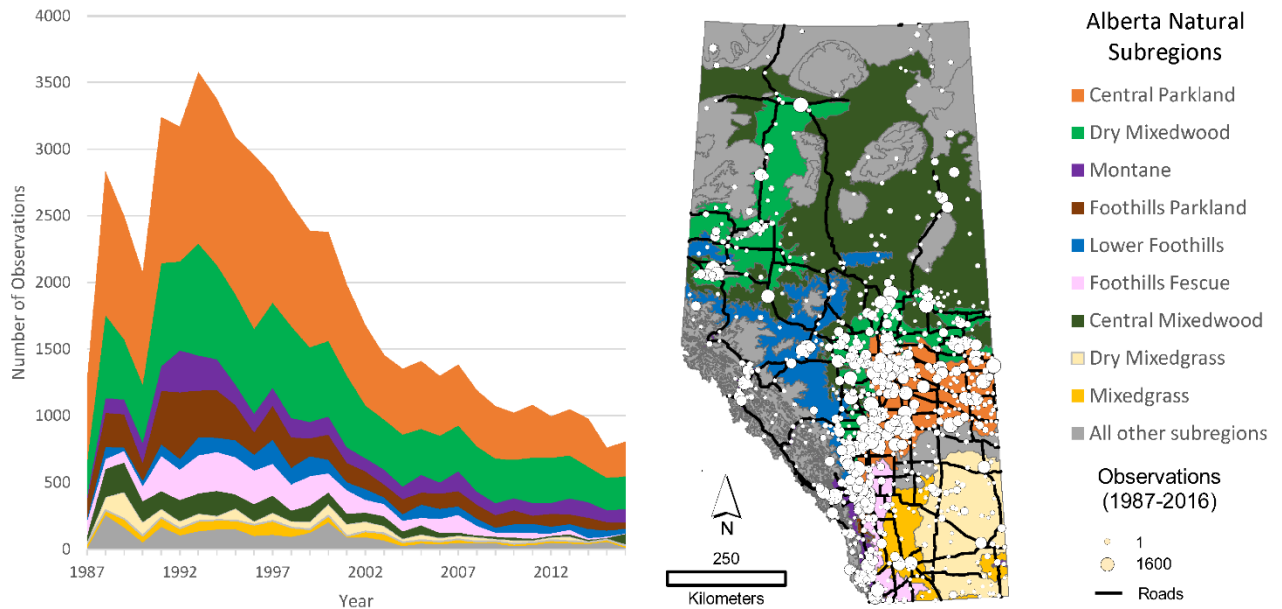


**Figure 2.** Natural subregions for ecological classifications used for data cleaning Method 1 Standardized difference, Method 3 Linear model by natural subregion, and Method 4 Observer correlation.

## 2.2 Alberta PlantWatch phenological data

Phenological data from 1987 to 2016 was provided for this study by Elisabeth Beaubien, Alberta coordinator the citizen science phenological program PlantWatch. Data was checked at the manual entry stage for easily recognized issues such as location, phenophase (phase), and species numbers that do not exist. These “impossible” records were cleaned manually, and cross referenced with the original data submission files to ensure errors resulting from data entry were limited as much as possible. In total, 57 745 observations were available for 30 species from 1987 to 2016. Observer locations are primarily where human population is greater i.e. within

areas around large cities and extensive road networks. Observations in remote areas are limited, such as those far away from road networks in the northeast portion of the province (Figure 3). Basic summary statistics for each plant species and phase used in data cleaning are available in Appendix A.



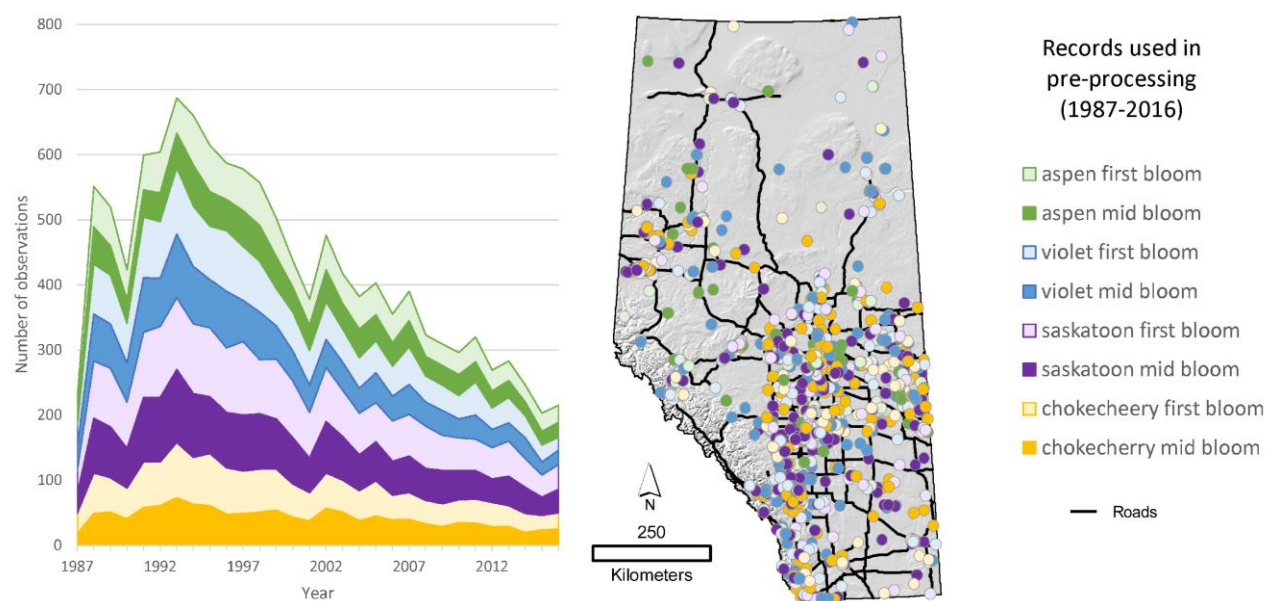
**Figure 3.** Observations made by Citizen Scientists for the Alberta PlantWatch program from 1987 to 2016. Observation counts and natural subregions are depicted with matching colours.

Instructions for phase reporting in the PlantWatch program are detailed in a “PlantWatch Canada in Bloom!” publication by Nature Canada (2010) and on the web ([www.plantwatch.ca](http://www.plantwatch.ca)). Phase 1 (first bloom) is reported as when the first flowers are open, or when male catkins or cones first start shedding pollen. When flowering occurs in multiple places across the plant, observers are instructed to report the date where first flowering or shedding is observed in three places across the plant. Phase 2 (mid bloom) is reported as when either 50% of flowers are open, or when 50% of male catkins or cones are shedding pollen. Additional phases can be reported by observers for full bloom and leafing however these were not consistently reported for all species’ and across all years as protocols were adjusted over the course of the study period (Beaubien and



Hamann 2011b). As a result, only first bloom and mid bloom phase data was used for the purposes of this research project.

Of species data collected over the whole study period, four species were selected where a minimum of 18 records for first bloom and mid bloom phases were recorded from 1987 to 2016. The four species used for this study included the tree aspen (*Populus tremuloides* Michx.), the herbaceous species early blue violet (*Viola adunca* J.E.Smith), and the two shrub species chokecherry (*Prunus virginiana* L.) and saskatoon (*Amelanchier alnifolia* Nutt.) (Figure 4).



**Figure 4.** Observations made for the Alberta PlantWatch program from 1987 to 2016 used as a part of the data cleaning study. Observer counts and locations are depicted with matching colours.

### 2.3 Climatic data

In this study, 1 km x 1 km gridded climatic data was obtained from the National Aeronautics and Space Administration, Daymet project for each observation point (Thornton et al, 2016; Hufkens 2017). The climatic variables were summed from January 1 up to the day of year that each phenological observation was made, and included the cumulative maximum daily temperature, cumulative minimum daily temperature, cumulative average daily temperature

(calculated by the average of the maximum and minimum temperatures), cumulative daily day length, cumulative daily precipitation, cumulative daily solar radiation, cumulative daily snow water equivalent, and cumulative daily water vapor pressure.

## **2.4 Data cleaning methods used for phenological data quality improvement**

Data cleaning methods used for improvement of quality for phenological data was as follows: percentile-based outlier removal, residuals of linear models, correlations among observers, and deviations from multivariate clusters.

Each species and phase (species phase) was cleaned independently of other species phases. In total, 5% of data points were removed with each data cleaning method. Data cleaning activities, summary statistics, and visualizations were done using the R programming environment (R Development Core Team 2014). These five methods are introduced below in more detail.

### *Method 1 - Standardized difference (SDiff)*

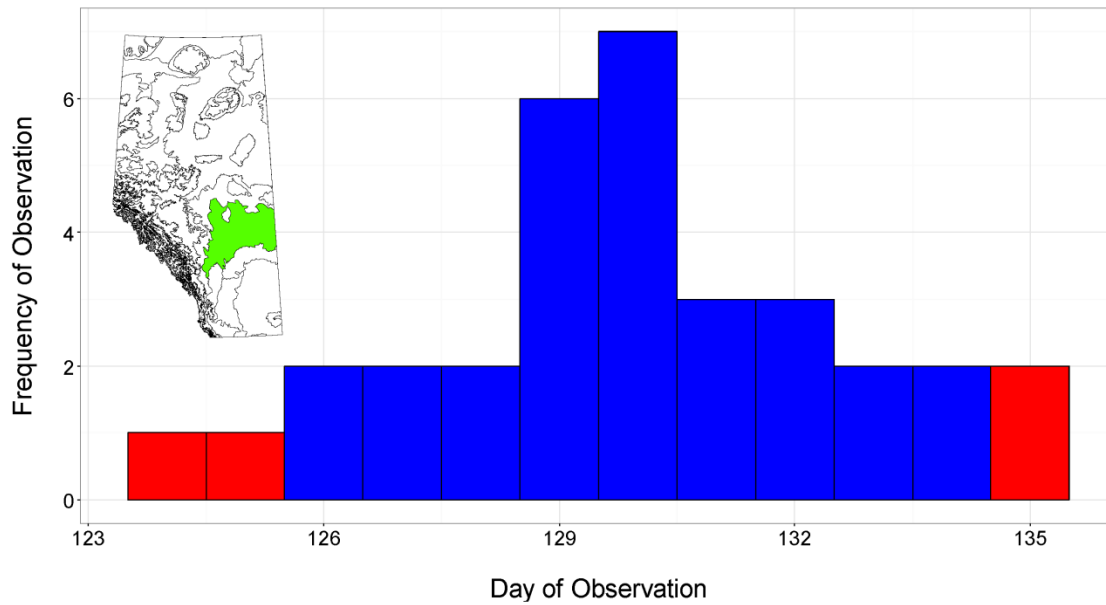
Natural subregions generally have similar climatic factors, landscape pattern, vegetation, soils, elevation, and physiographic features (Natural Regions Committee 2006). Native plants and trees have adapted to the local factors (Savolainen et al 2007). Thus, the expectation is that observations of the same species phase are expected to occur at approximately the same time within nearby areas, and within individual natural subregions. A good indication for a potential error would therefore be the tails of the distribution around the mean value of the natural subregion (Figure 5).

In order to compare outliers across different natural subregions, the scale of observations was standardized to express each observation in units of standard deviations from the natural subregion mean using Cohen's d, grouped by natural subregion and year (subregion year):

$$S_{diff} = \frac{x - \bar{x}}{s} \quad (1)$$

Where  $S_{diff}$  is the values scaled to the standard deviation of the mean,  $x$  is the day of observation,  $\bar{x}$  is the average observation date grouped by subregion year, and where  $s$  is the standard deviation of observations.

Where only one record for a subregion year is available, the standard deviation is set to zero. 5% of observations were removed with the highest standardized difference from the mean of their corresponding subregion year.



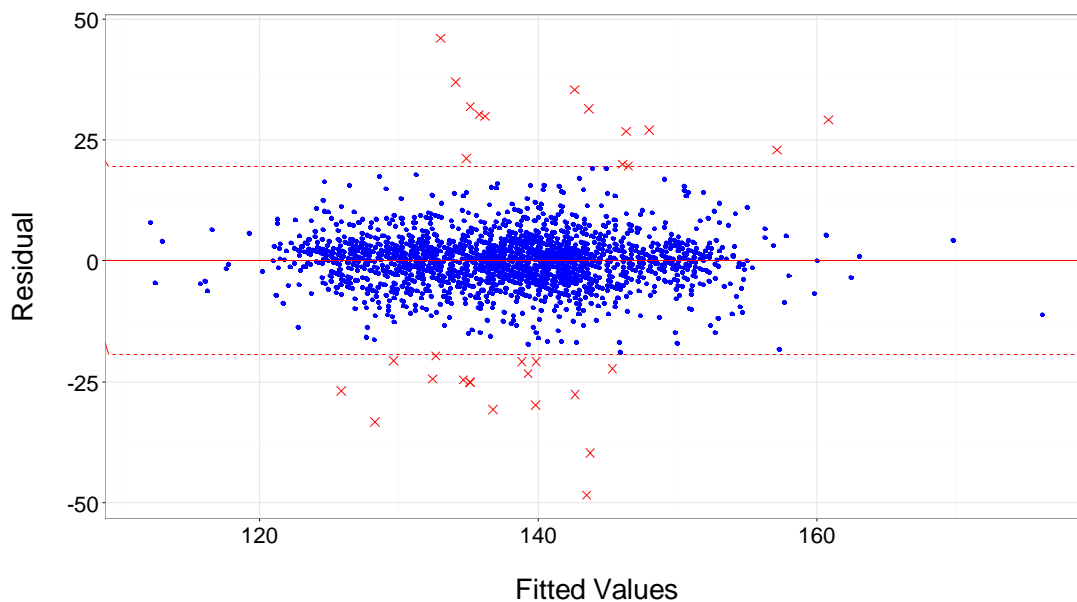
**Figure 5.** Conceptual depiction of Method 1 (Standardized difference) data cleaning based on the removal of outliers grouped by natural subregion and year. Blue represents conceptually retained records, red represents conceptually removed records.

*Method 2 - Linear model (LM)*

This method of data cleaning is based on developing a generalized linear model for the purposes of data cleaning (Figure 6). This model is similar to the model developed by Ranjitkar (2013), where flowering is based on a linear model developed using location (latitude, longitude, and elevation) as predictors. Since the model developed by Ranjitkar (2013) was for observations

made over the course of one year, an additional predictor was included in this study in order to account for annual climatic variability such as El Niño and El Niña events.

For the development of this (LM) method of data cleaning, latitude, longitude, elevation, as well as year were used as predictors in the linear model for the phenological occurrence. The records for removal were assessed by calculating the residual difference between the predicted day of year of phenological occurrence (based on the linear model), and the observed day of occurrence. The 5% of observations with the highest absolute residual difference to the linear model were removed.



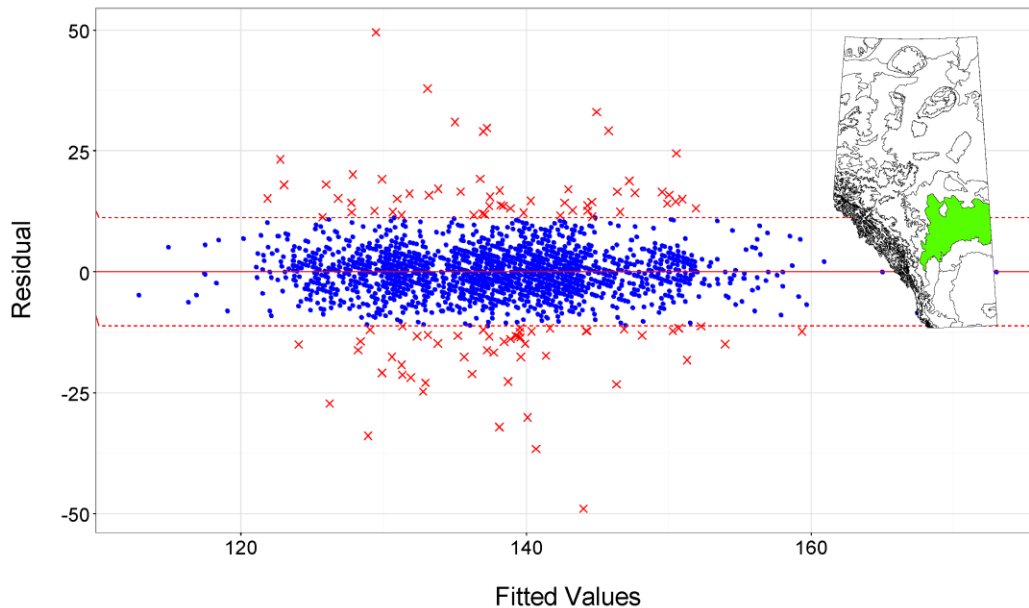
**Figure 6.** Conceptual depiction of Method 2 (Linear model) data cleaning, based on the development of a linear model with latitude, longitude, elevation, and year as predictors. Observations with the largest residual difference are removed. Blue represents conceptually retained records, red 'x' represents conceptually removed records.

### *Method 3 - Linear model by natural subregion (LMNSR)*

As outlined in Method 1 (SDiff) above, the expectation is for observations of the same species phase to occur at approximately the same time within individual natural subregions. As a modification of Method 2 (LM) above, Method 3 (LMNSR) uses natural subregion and year as predictors for developing a linear model for the removal of outliers (Figure 7). By using natural

subregions, local variability is accounted for, such as landscape pattern, climatic variables, and elevation, as these factors were incorporated into delineation of natural subregions (Natural Regions Committee 2006).

The residual difference was again calculated by subtracting the predicted day of year of phenological occurrence (based on the LMNSR linear model) from the observed day of occurrence, and the 5% of observations with the highest absolute difference were removed.

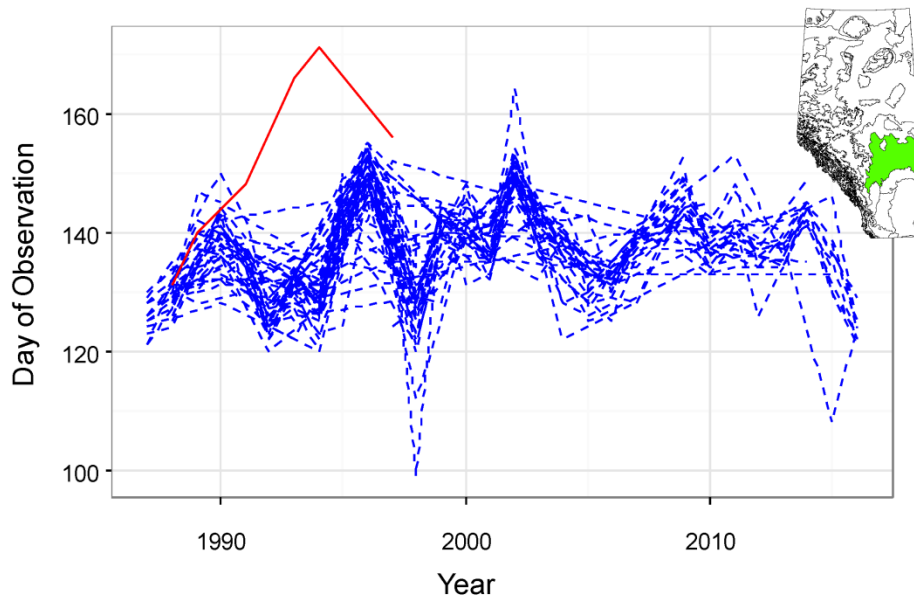


**Figure 7.** Conceptual depiction of Method 3 (Linear model by natural subregion) data cleaning, based on the development of a linear model with natural subregion and year as predictors. Observations with the largest residual difference are removed. Blue represents conceptually retained records, red 'x' represents conceptually removed records.

#### *Method 4 - Observer correlation (CORR)*

In a previous study using Alberta PlantWatch data, it was found that one-time observers are relatively unbiased in over and under-estimation of observations compared to long-term observers (Beaubien and Hamann 2011b). This was found although the quality of observations can be influenced by training of observers (Kosmala et al 2016). Method 4 (CORR) is based on the idea that by removing potentially inconsistent or unreliable observers, the accuracy of the dataset may be improved. Since observations of the same species phase are expected to occur at

approximately the same time within individual natural subregions, an individual's observation within a subregion year should correlate with the overall mean of all observers' phenological activity within the natural subregion as well (Figure 8). By limiting observations within an individual natural subregion, this method also takes into account local variability. While observers are instructed to submit only one date for each species phase a year, there were a few observers that submitted multiple observations for the same species phase within the same subregion year. As a result, prior to calculating the correlation of observers, the day of year for phenological events was averaged for each observer where the species, phase, natural subregion, and year all matched.



**Figure 8.** Conceptual depiction of Method 4 (Observer correlation) data cleaning, based on the removal of observers with the lowest correlation to the average observations by natural subregion. Blue dashed lines represents conceptually retained observers; red solid line represents a conceptually removed observer.

Where observers submitted three or more observations for an individual species phase within a natural subregion over several years, a correlation was calculated for the observer's values to the overall mean of the natural subregion, and ranked from lowest to highest correlation values by observer. Correlation was calculated using the 'cor' function in R with the default Pearson

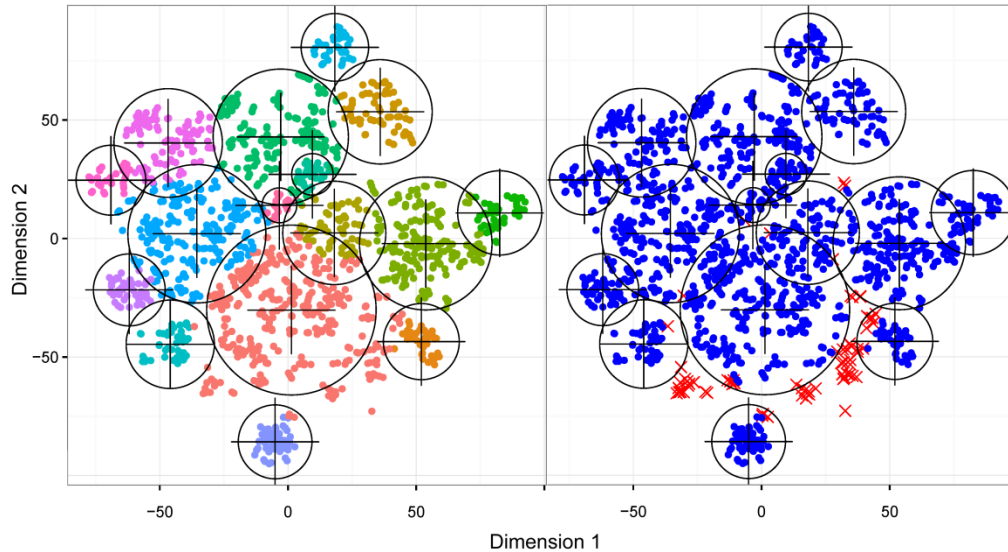
correlation coefficient settings. If observers have submitted only one or two total observations over all years for a natural subregion species phase, correlation to the mean could not be determined. These observers were ranked by comparing the absolute difference of their individual observations from the mean of the subregion year in the same process as Method 1 (SDiff) above, using equation (1).

Observers in the correlation dataset (with three or more observations per subregion) with the lowest correlation to the mean species phase values, and observers with observations having the highest residual difference in the standardized difference dataset (with 1 or 2 observations) were noted and all records from those observers in the natural subregion were removed from the dataset. An attempt was made to remove 5% of observations from the correlation dataset (with 3 or more observations) and 5% from the standardized difference dataset (with 1 or 2 observations). However, since only whole observers were removed, some variation occurred in the proportion of observers in each dataset in order to meet the 5% target removal. This method is perhaps the most subjective in terms of observation removal. This is because, in order to have an exact 5% removal, the selection of observers could not be automated.

#### *Method 5 Dimensionality reduction and clustering (DRC)*

Previous studies have used spatiotemporal proximity to verify observations. However, this (DRC) method of data cleaning incorporates not only geographic proximity, but also environmental contextual information to explain inconsistencies in information (Mehdipoor et al 2015).

This dimensionality reduction and clustering (DRC) method is based on a workflow devised by Mehdipoor et al (2015) for phenological data (Figure 9). Although Mehdipoor et al (2015) used USANPN data in developing their workflow, it may also be used for Alberta PlantWatch data. This is since rather than using the sequence yes/no observations for observation date, the Mehdipoor et al (2015) workflow uses the day of year of a given phenological phase (flowering onset) as the observation date, similar to the Alberta PlantWatch program protocol.



**Figure 9.** Conceptual depiction of Method 5 (Dimensionality reduction and clustering) data removal method. T-distributed stochastic neighbor embedding dimensionality reduction reduces data to two dimensions and clustered into groups (left panel). The points with the largest standardized difference to its cluster mean are removed. Blue represents conceptually retained records, and red 'x' represents conceptually removed records (right panel).

The day of year of the phenological observation, latitude, longitude, elevation, and climatic variables obtained through the National Aeronautics and Space Administration Daymet project (Thornton et al 2016; Hufkens K 2017) were used as into the tsne package in R for dimensionality reduction (Donaldson 2012). T-distributed stochastic neighbor embedding (t-SNE) was used as a means for visualizing high dimensional data in low dimensional space. The t-SNE, which was originally developed by van der Maaten and Hinton (2008), allows for both separation of dissimilar data points, as well as preserving short distances of similar datapoints. Perplexity settings ranged from 5 to 50 with increments of 5. This also followed the workflow outlined by Mehdipoor et al (2015). In total, 5000 iterations were run using the tsne package in R in order to allow sufficient iterations for the error value to stabilize.

Clustering was conducted by optimizing the Bayesian Information Criterion (BIC) within the mclust package in R (Fraley et al 2012). Following the workflow outlined by Mehdipoor et al (2015), this study automated the clustering method selection. The BIC was calculated for up to 100 clusters, and the clustering model with the highest BIC was automatically selected for each perplexity value.



Outliers were removed for all perplexities by first calculating the Euclidean distance of each observation from the center of its respective cluster. The standardized difference was calculated for each data point to the center of the cluster using Cohen’s d (equation 1). In total, 5% of data points with the highest standardized distance to their respective cluster center were removed.

After dimensionality and reduction was run on all perplexities, Moran’s I was calculated for each year of all perplexity values, and the perplexity value that resulted in the greatest average increase in Moran’s I value was selected as the overall “best” perplexity for each species phase.

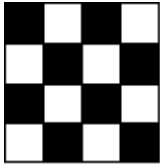
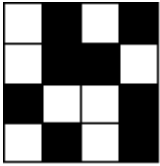
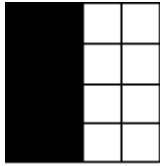
## 2.5 Assessment of validity for different data cleaning methods

The degree of improvement in data cleaning provided by the five cleaning methods was assessed using spatial autocorrelation, quantified by Moran’s I. Moran’s I, which ranges from -1 to 1, is comparable to a correlation coefficient, where 0 indicates random patterns depending on the strength of the autocorrelation (Latta et al 2009), positive value’s for Moran’s I indicate positive spatial autocorrelation, and negative values indicate negative spatial autocorrelation (Table 1). For all data cleaning methods, Moran’s I was calculated using the ape package in R (Paradis et al 2004) with an inverse distance matrix. Records were grouped by year, and species phase for calculation. Moran’s I was utilized rather than the probability factor since the probability factor can reflect the significance of negative or positive autocorrelation. The equation used to calculate Moran’s I is:

$$I = \frac{n \sum_i^n \sum_j^n w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_i^n \sum_j^n w_{ij} \sum_i^n (X - \bar{X})^2} \quad (2)$$

Where I is Moran’s I, n is the number of observations of variable x, at locations i, j.  $\bar{x}$  represents the mean of the x variable over n locations, and  $w_{ij}$  represents the spatial weights matrix for locations i and j (Zhou and Lin 2008).

**Table 1.** Conceptual examples of spatial autocorrelation

	Negative Spatial Autocorrelation	Random Spatial Pattern	Positive Spatial Autocorrelation																																																
Illustrative example																																																			
Numerical example	<table border="1" data-bbox="488 569 646 699"> <tr><td>5</td><td>1</td><td>7</td><td>2</td></tr> <tr><td>1</td><td>6</td><td>3</td><td>6</td></tr> <tr><td>7</td><td>1</td><td>7</td><td>2</td></tr> <tr><td>1</td><td>7</td><td>1</td><td>6</td></tr> </table>	5	1	7	2	1	6	3	6	7	1	7	2	1	7	1	6	<table border="1" data-bbox="764 569 922 699"> <tr><td>1</td><td>7</td><td>1</td><td>7</td></tr> <tr><td>2</td><td>6</td><td>5</td><td>2</td></tr> <tr><td>7</td><td>2</td><td>3</td><td>6</td></tr> <tr><td>1</td><td>5</td><td>2</td><td>7</td></tr> </table>	1	7	1	7	2	6	5	2	7	2	3	6	1	5	2	7	<table border="1" data-bbox="1084 569 1242 699"> <tr><td>1</td><td>2</td><td>6</td><td>6</td></tr> <tr><td>1</td><td>3</td><td>7</td><td>6</td></tr> <tr><td>2</td><td>1</td><td>5</td><td>7</td></tr> <tr><td>1</td><td>2</td><td>6</td><td>5</td></tr> </table>	1	2	6	6	1	3	7	6	2	1	5	7	1	2	6	5
5	1	7	2																																																
1	6	3	6																																																
7	1	7	2																																																
1	7	1	6																																																
1	7	1	7																																																
2	6	5	2																																																
7	2	3	6																																																
1	5	2	7																																																
1	2	6	6																																																
1	3	7	6																																																
2	1	5	7																																																
1	2	6	5																																																
Moran's I Statistic <sup>1</sup>	Negative, approaching -1	Positive or negative, approaching 0	Positive, approaching 1																																																

<sup>1</sup> Values vary based on the weight matrix used in calculation of Moran's I statistic.

Since Moran's I calculation outlined in equation (2) incorporates the number of records in a dataset in its calculation, the number of records removed per species phase was consistent for all cleaning methods. While in this study 5% removal is set in order to remove equal proportions of records, this removal limit may be adjusted by end users based on required geographic extent and research goals.

Moran's I was calculated for each year (1987 to 2016) for each of the eight species phases selected for this study, and for each of the five data cleaning methods as well as the original datasets prior to cleaning. These values were evaluated using a two way ANOVA model with the year and species phase as random effects.

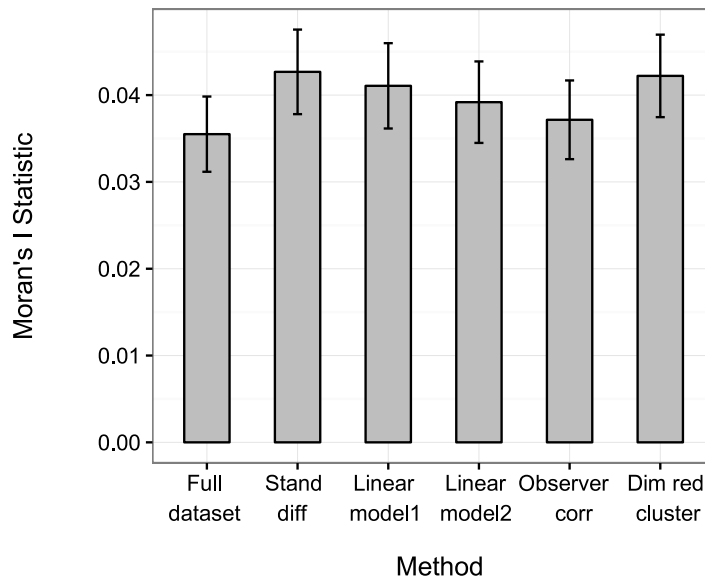
The full range of records before data cleaning, and with each data cleaning method were plotted using the ggplot2 package in R (Wickham 2009) and visually assessed for variability in the data (Figure 11). Outliers were classified as having values beyond the interquartile range (the 25<sup>th</sup> percentile and 75<sup>th</sup> percentile) multiplied by 1.5. The day of phenological occurrence for each natural subregion was predicted using a Best Linear Unbiased Prediction model with the asreml package in R (Butler 2009). In this model, the phase was set as the predictor, and year, species,

and natural subregion were random effects. Records that were removed with each data cleaning method were also mapped and visually assessed for spatial trends.

### 3. RESULTS

#### 3.1 Improvement of overall data quality

All data cleaning methods demonstrated an increase in Moran's I when compared to the Moran's I generated with the full datasets prior to data cleaning (Figure 10). The two methods with the greatest increase in Moran's I value were Method 1 (SDiff) and Method 5 (DRC) and had only a small difference in their Moran's I values. The improvement in Moran's I was the highest with Method 1 (SDiff) and Method 5 (DRC) followed by Method 2 (LM), Method 3 (LMNSR), and Method 4 (CORR) respectively. However, the differences for Moran's I value between the five data cleaning methods were not statistically significant (Table 2). Summary statistics for Moran's I values are provided in Appendix B.



**Figure 10.** Effectiveness of data cleaning methods as measured by Moran's I statistic ( $\pm$ SE) before and after data cleaning. N=240 groups of datasets per cleaning method (4 species x 2 phases x 30 years). 5% of records were removed during data cleaning for each data cleaning method. The "full dataset" represents the Moran's I statistic for the original records prior to data cleaning.

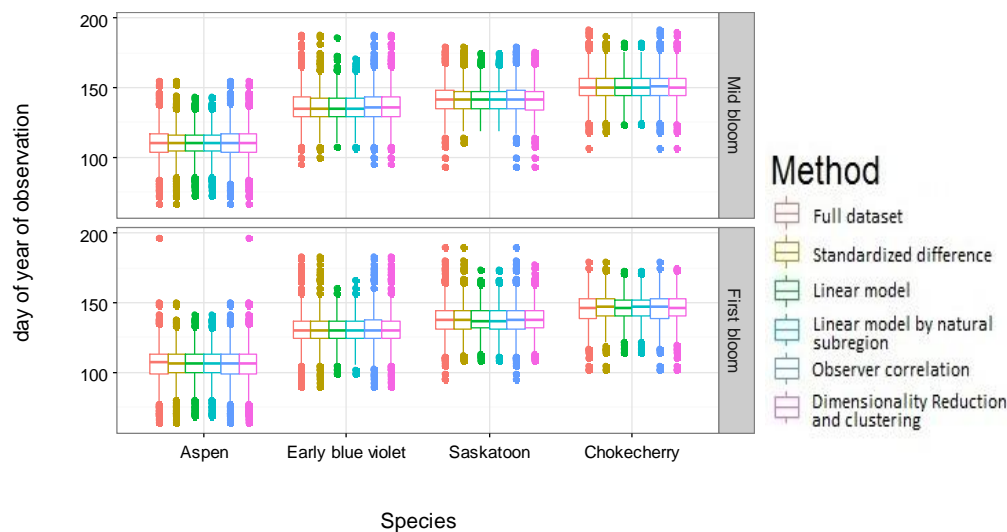
**Table 2.** Two-way analysis of variance (ANOVA) results of Moran’s I statistic values, with individual year (1987-2016) and species phase (4 species x 2 phases) as random effects. Method factors include the original dataset as well as five assessed methods of data cleaning (5% of records were removed during data cleaning).

Source	DF	SS	MS	F	P
Method	5	0.0099	0.001986	0.3955	0.852
Year	1	0.3070	0.306999	61.1516	1.018x10 <sup>-14</sup> ***
Species phase	1	0.0641	0.064102	12.7687	3.642 x10 <sup>-4</sup> ***
Total	1432	7.1891	0.005020		

\*\*\* denotes significance at p<0.001.

### 3.2 Global outlier treatment

Based on visual assessment, Method 1 (SDiff) and Method 4 (CORR) both retained many of the prevalent outliers resulting in a greater range of variability in the resulting (cleaned) data. Method 2 (LM) and Method 3 (LMNSR) both removed many of the prevalent outliers, resulting in a more apparent reduction in the range and variability of the data. Method 5 (DRC) did not appear to follow any consistent trends in the variability in the resulting (cleaned) data. This is visible in first bloom of aspen, where an abnormally late bloom was retained, especially compared to first bloom of saskatoon, where many of the late outliers were removed (Figure 11). More detailed summary statistics for original (raw) dataset and cleaned data (after outlier removal) are provided in Appendix A.



**Figure 11.** Differences among data cleaning methods in the treatment of global outliers for five assessed data cleaning methods. Outliers were classified as having values beyond the interquartile range (the 25th percentile and 75th percentile) multiplied by 1.5.

### 3.3 Effects of data cleaning on regional means

While means and medians across the entire dataset and all years were not affected at all by data cleaning methods (Figure 11), we expect changes to become more pronounced when separately analyzed for different natural subregions, years, and species. Nonetheless, for many of the species-region-year combinations phenology estimates show only minor shifts in dates after the five data cleaning methods have been applied. Ninety-five percent of all mean dates for species-region-year combination shift by one to four days, depending on the cleaning method (Table 3).

**Table 3.** Change in predicted date for phenological occurrence date for natural subregions after data cleaning compared to the original predicted date for phenological occurrence.

Method	Maximum change (days) <sup>1</sup>		95 <sup>th</sup> percentile of magnitude of change (days) <sup>2</sup>	Change in standard error of the estimate compared to control (days)
	Earliest	Latest		
Method 1 (Standardized difference)	-6.6	4.7	±1.5	0.29
Method 2 (Linear model)	-13.1	15.2	±4.4	-0.92
Method 3 (Linear model by natural subregion)	-13.0	14.7	±4.0	-0.84
Method 4 (Observer correlation)	-4.3	5.9	±1.1	0.16
Method 5 (Dimensionality reduction and clustering)	-6.2	6.7	±2.0	0.16

<sup>1</sup>Maximum change represents the largest changes observed in predicted phenological occurrence date. Best Linear Unbiased Prediction was used to predict the phenological occurrence date in each natural subregion using the asreml package in R (Butler 2009). Phase was the predictor, and year, species, and natural subregion were random effects.

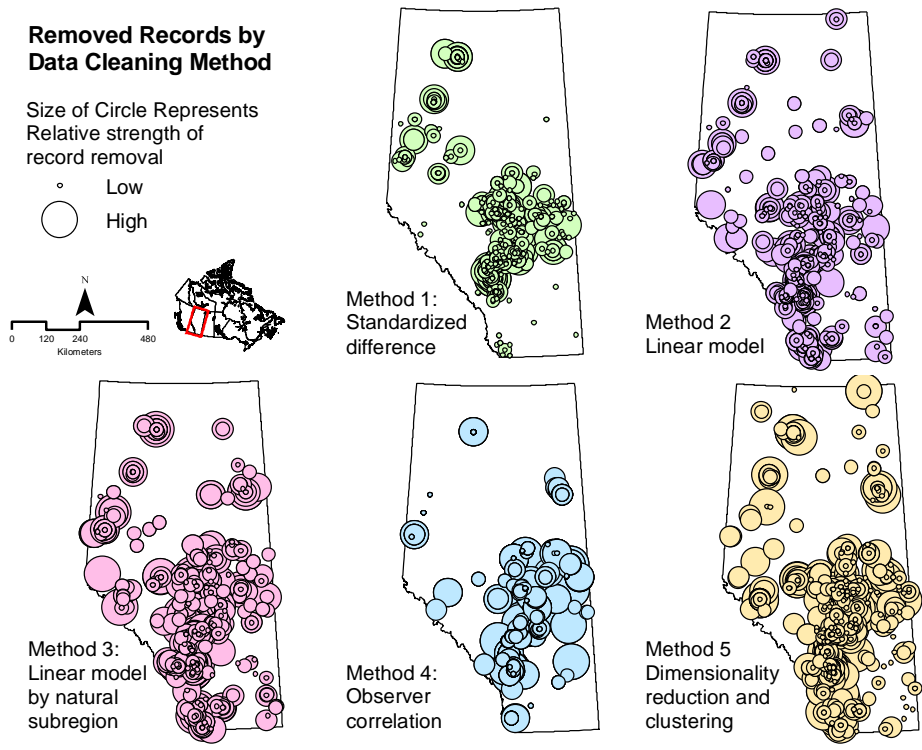
<sup>2</sup>Difference in days to the mean, calculated using the standard deviation of the differences between the original and post-cleaning occurrence date, multiplied by 2.

Predicted phenological occurrence date within natural subregions shifted by the greatest number of days (both positive and negative) with the two linear models Method 2 (LM) and Method 3 (LMNSR), i.e. being up to 15.2 days later than the original predicted date. The variation of the change in the predicted day of phenological occurrence was also greatest with the linear model. The 95% of all predicted occurrences after data cleaning were within up to  $\pm 4.4$  days difference compared to the original natural subregion prediction. However, with the linear model, the standard error of the phenological occurrences within natural subregions decreased with the two linear data cleaning methods by up to 0.92 days.

The shifts in predicted date for phenological occurrence in each natural subregion did not vary as much with any of the other methods of data cleaning (SDiff, CORR, and DRC). Maximum change in predicted day of phenological occurrence for a natural subregion ranged from 4.3 to 6.7 days earlier or later than the original predicted phenological occurrence date, and the confidence interval was narrower with 95% of post-cleaning estimates being within  $\pm 2$  days of the original mean. However, these three different data cleaning methods demonstrated an increase in the standard error of phenological occurrence within natural subregions.

### **3.4 Spatial patterns for removal of records due to data cleaning**

Records that were removed with each data cleaning method were also mapped for spatial trends (Figure 12). While all data cleaning methods removed records in high density areas (in the south-central region of Alberta), the influence of data cleaning methods varied in low density areas, in particular the northeast and southeast portions of the province.



**Figure 12.** Spatial locations of removed records for assessed data cleaning methods. For Methods 1 and 5, the size of the circle represents the absolute difference from the mean. For Methods 2 and 3, the size of the circle represents the absolute difference from the model, and for Method 4, the size of the circle represents the correlation of removed observers multiplied by -1.

Method 1 (SDiff) and Method 4 (CORR) appear to preferentially remove records in high density observation areas and retain records in low density areas when compared to any of the other cleaning methods. By comparison, Method 5 (DRC) appeared to preferentially remove records in the low density observation areas of the northeast portion of the province. Methods 2 (LM), 3 (LMNSR), and 4 (CORR) did not appear to preferentially remove records in high or low density areas. However, a few more records in the northeast portion of the province were retained with Method 2 (LM) when compared to Method 3 (LMNSR). Removed records through Method 4 (CORR) also appeared to be more clustered than any of the other data cleaning methods.

### 3.5 Similarity of outlier rankings between data cleaning methods

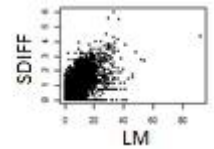
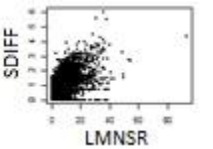
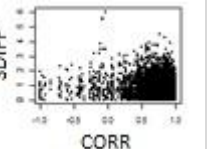
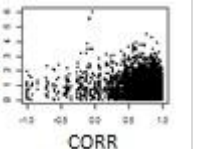
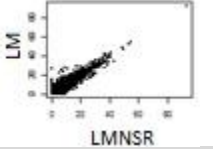
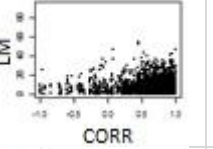
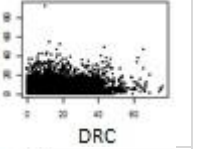
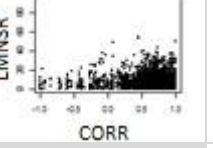
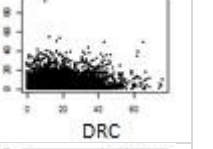
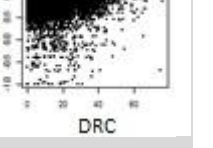
Since the number of records removed was kept consistent, 638 records were removed across all species phases for each cleaning method. Of these, only 13 removed records were consistent across all cleaning method (2%). Ten of these records were for saskatoon (six first bloom, four mid bloom), one record was for violet (first bloom), one record was for aspen (mid bloom) and one record was for chokecherry (first bloom). Three of these thirteen records were recorded in 1994, two in 1993, 1998, and 1999 respectively, and one in 1988, 1997, 2011, and 2015. While the majority of these consistently removed records were for saskatoon (10 of the 13, or 77%), other prevalent patterns were not observed in the phase or year of consistently removed records.

Despite the low number of consistently removed records, all data cleaning methods demonstrated a significant correlation to other methods, with varying r-values (Table 4). With Method 4 (CORR), since the correlation value is based on removal of low or negative values, a negative correlation confirms a relationship between point removal with the CORR method, and with all other data cleaning methods.

Method 1 (SDiff), Method 2 (LM), and Method 3 (LMNSR) all demonstrated relatively strong positive r-values ranging from 0.58 (between LM and SDiff  $p < 0.05$ ) to 0.93 (between LM and LMNSR  $p < 0.05$ ). A weak r-value was present with CORR where r-values ranged from -0.1558 (with LM  $p < 0.05$ ) to -0.1779 (with SDiff  $p < 0.05$ ). The weakest r-value was present with DRC where r-values ranged from 0.0391 (with SDiff  $p < 0.05$ ) to 0.1251 (with LM  $p < 0.05$ ).



**Table 4.** Illustration of correlations, and statistical strengths of correlations between removal values amongst five assessed data cleaning methods Method 1 Standardized Difference (SDiff), Method 2 Linear model (LM), Method 3 Linear model by natural subregion (LMNSR), Method 4 Observer correlation (CORR), and Method 5 Dimensionality reduction and clustering (DRC) (method = Pearson).

	SDiff <sup>A</sup>	LM <sup>B</sup>	LMNSR <sup>B</sup>	CORR <sup>C</sup>	DRC <sup>A</sup>
SDiff					
LM	r: 0.5788 p: < 2.2e-16				
LMNSR	r: 0.6115 p: < 2.2e-16	r: 0.9313 p: < 2.2e-16			
CORR	r: -0.1779 p: < 2.2e-16	r: -0.1558 p: < 2.2e-16	r: -0.1673 p: < 2.2e-16		
DRC	r: 0.0391 p: 9.8e-6	r: 0.1251 p: < 2.2e-16	r: 0.1208 p: < 2.2e-16	r: -0.0535 p: 4.2e-8	

<sup>A</sup> Correlation values based on the standardized difference of an observation to the mean value of respective groupings.

<sup>B</sup> Correlation values based on the absolute difference of an observation to a linear model.

<sup>C</sup> Correlation values based on the correlation of the individual observer (where 3 or more observations per species phase within an NSR has been submitted) to the overall mean.

### 3.6 Other findings for specific methods

Method 4 (CORR) used a mixed approach to data cleaning where correlations to the overall mean were calculated for individual observers with more than 3 submitted observations for a specific species phase within any given natural subregion. A standardized difference approach (calculated with Cohen's d given in equation 1) was used to rank observations when only one or two observations for a species phase within any given natural subregion were reported for an individual observer in the dataset.

Through this method, 505 of the 638 records (79%) of the records were removed through correlation and the remaining 133 records (21%) were removed through the standardized difference approach (Table 5).

Method 5 (DRC) followed the workflow outlined by Mehdipour et al (2015). However, rather than defining a user-selected perplexity for the dimensionality reduction process, the perplexity selection was automated by selecting the perplexity level with the highest increase in Moran’s I. The perplexities and the clustering methods were not consistent across species phases (Table 6).

When assessing the clusters for each species phase individually, observations within the same cluster were generally spatially clustered. However, in south central regions of the province, where high densities of observations are present, there was considerable and visible overlap in clusters (Figure 13). This data cleaning method required extensive computational time when compared to the other data cleaning methods. The runtime for the dimensionality reduction and clustering scripts for each species phase, and for all perplexities assessed required approximately 8 hours, but may vary depending on computational capacity

**Table 5.** Proportion of removed records through Method 4 (Observer correlation) data cleaning. A mixed approach to data cleaning was used based on the number of records submitted by any individual observer within any given natural subregion. In total, 5% of the overall records were removed for each species phase.

	Correlation removal (with 3 or more records) <sup>A</sup>		Standardized difference removal (with 1 or 2 records) <sup>B</sup>	
	No. of observers removed	No. of records removed	No. of observers removed	No. of records removed
Aspen First bloom	11 of 152 (7.2%)	55 of 1166 (4.7%)	13 of 200 (6.5%)	15 of 234 (6.4%)
Aspen Mid bloom	8 of 127 (6.3%)	48 of 968 (5.0%)	7 of 176 (4.0%)	11 of 215 (5.1%)
Early Blue Violet First bloom	15 of 186 (8.1%)	64 of 1437 (4.4%)	17 of 255 (6.7%)	23 of 317 (7.3%)
Early Blue Violet Mid bloom	16 of 173 (9.2%)	59 of 1228 (4.8%)	13 of 366 (3.6%)	18 of 311 (5.8%)
Saskatoon First bloom	15 of 211 (7.1%)	88 of 1841 (4.8%)	16 of 287 (5.6%)	21 of 358 (5.9%)
Saskatoon Mid bloom	19 of 209 (9.1%)	79 of 1599 (4.9%)	15 of 280 (5.4%)	18 of 348 (5.2%)
Chokecherry First bloom	14 of 150 (9.3%)	60 of 1207 (5.0%)	11 of 226 (4.9%)	14 of 288 (4.9%)
Chokecherry Mid bloom	11 of 136 (8.1%)	52 of 1033 (5.0%)	9 of 198 (4.5%)	13 of 252 (5.2%)

<sup>A</sup> correlations to the overall mean were calculated for individual observers when more than 3 submitted observations for a specific species phase within any given natural subregion were present in the dataset.

<sup>B</sup> observations were ranked through standardized difference (Cohen’s d) when only one or two observations for a species phase within any given natural subregion were reported for an individual observer in the dataset.

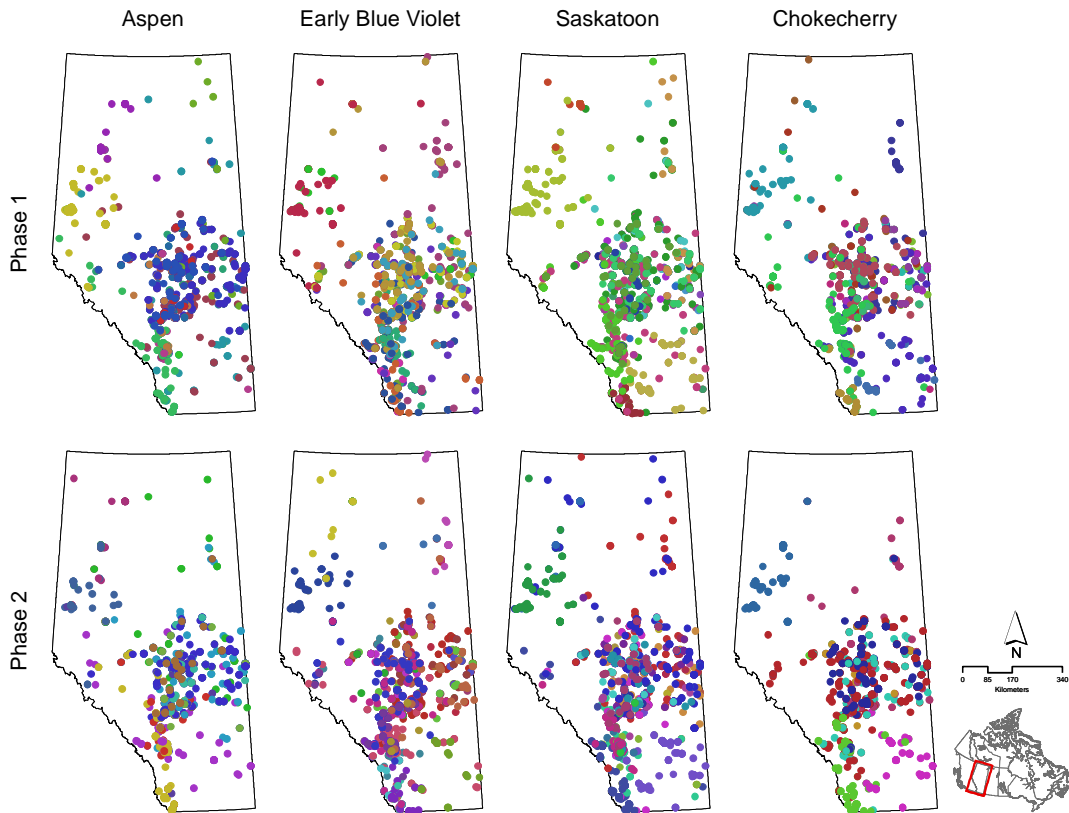
**Table 6.** Fitted dimensionality reduction and clustering models from Method 5: Dimensionality reduction and clustering technique for automated data cleaning.<sup>A</sup>

	Perplexity level <sup>B</sup>	Error at 5000 iterations <sup>B</sup>	BIC value <sup>C</sup>	Clustering method <sup>C</sup>	No. of clusters <sup>C</sup>
Aspen First bloom	45	0.774	-27582.4	spherical, unequal volume	14
Aspen Mid bloom	35	0.710	-24368.0	diagonal, varying volume, equal shape	12
Early Blue Violet First bloom	35	0.815	-35780.8	ellipsoidal, equal shape	12
Early Blue Violet Mid bloom	20	0.813	-33332.9	diagonal, varying volume, equal shape	25
Saskatoon First bloom	40	0.751	-45051.7	ellipsoidal, equal shape and orientation	19
Saskatoon Mid bloom	45	0.770	-39553.7	ellipsoidal, equal orientation	23
Chokecherry First bloom	40	0.760	-29878.9	ellipsoidal, equal shape	15
Chokecherry Mid bloom	45	0.763	-26148.0	ellipsoidal, equal shape	8

<sup>A</sup> Methodology modified from the workflow presented by Mehdipoor et al (2015)

<sup>B</sup> Automated dimensionality reduction was conducted using the tsne package in R

<sup>C</sup> Automated clustering was conducted using the mclust package in R by maximizing the Bayesian Information Criterion (BIC) value.



**Figure 13.** Automated clusters generated through Method 5 (Dimensionality reduction and clustering). Climatic variables, location, and observed phenological occurrence date were incorporated for automated dimensionality reduction and clustering. Each colour represents and individual cluster grouping.

#### 4. DISCUSSION

Potential sources of human-caused inconsistencies in the Alberta PlantWatch dataset may be due to variation in the training protocol and effort expended by the observer. Additional errors that could be present in the dataset may be the result of incorrect identification of plants, and data entry errors – either at the time of observation, or at the time of manual database entry. The goal of removing inconsistent or potentially inaccurate observations is to improve the consistency of the dataset.

The overall mean and medians for full species phase datasets after data cleaning did not shift in this study by more than one day. In comparison, in the study by Mehdipoor et al (2015), the post-cleaned data shifted by two days per decade of data. However, when the predicted regional mean of phenological occurrence by natural subregion was estimated using a Best Linear Unbiased Prediction model, the predicted average regional bloom date shifted in this study by up to 15.2 days after data cleaning, and 95% of post-cleaning regional prediction dates were within  $\pm 4.4$  days of the original predicted date. The largest shifts in regional predicted means were for the two linear model methods of data cleaning (Method 2 LM and Method 3 LMNSR). These dramatic shifts in predicted natural subregion means using the two linear model methods indicate a large influence of prevalent outlier removal in the dataset (as illustrated in Figure 11), and larger influence of data cleaning on phenological occurrence trends.

The shifts in predicted regional mean for phenological occurrence was less dramatic with the other three methods of data cleaning, with 95% of post-cleaning predictions being within  $\pm 2$  days of the original regional mean. The maximum shift in predicted day of national subregion occurrence was 6.7 days after data cleaning with Method 5 (DRC), and diminishing with Method 1 (SDiff) and Method 4 (CORR).

Method 1 (SDiff) had the largest increase in Moran's I after data cleaning when compared to the Moran's I of the original dataset, followed by Method 5 (DRC), Method 2 (LM), Method 3 (LMNSR), and Method 4 (CORR) respectively. While Moran's I increased after data cleaning with all assessed methods, the change in Moran's I was not statistically significant for any data cleaning method.

The presence of positive r-values ranging from 0.58 to 0.93 between Method 1 (SDiff), Method 2 (LM), and Method 3 (LMNSR) indicate some consistency in the removal rankings of points between these three data cleaning methods. However, the absolute r-values for Method 4 (CORR) and Method 5 (DRC) were lower (less than 0.2). These lower consistencies in removal rankings of points for these two data cleaning methods indicate potentially different processes in treatment of data for the purposes of data cleaning. The weak correlation of Method 4 (CORR) to all other methods may be due to the removal of potentially accurate observations, through the removal of all records from that observer in the natural subregion. When comparing the r-values of Method 5 (DRC) to all other data cleaning methods, there was limited consistency between the removal rankings of records. All comparisons between removal rankings had absolute r-values less than 0.13.

A consideration for the use of each data cleaning method is the delineation of regions and the relative accuracy of delineations. Method 1 (SDiff), Method 3 (LMNSR), and Method 4 (CORR) incorporated the use of predefined natural subregions in their data cleaning algorithm. Natural subregion boundaries for Alberta were last updated over 10 years ago in 2006 using climatic observations from 1961 to 1990 (Natural Regions Committee 2006). For Method 1 (SDiff), a minimum of two records are required per natural subregion year, otherwise the standardized difference is set to zero. As a result, this method of data cleaning preferentially retains records within natural subregions with few observations. These may also include prevalent outliers and the observations collected in low density areas (illustrated in Figure 12). Similarly, Method 4 (CORR) was also limited to natural subregions in order to take into account local variability. The resulting (cleaned) data retained many prevalent outliers, and thus more variability in the data (illustrated in Figure 11).

While Method 2 (LM) did not require predefined regional delineation, this method of data cleaning inherently follows the assumption that by developing a linear model using location and year as the predictors, the progression of plant bloom is consistent across the province, and does not take into account local environmental factors such as landscape pattern, and local climatic trends. Method 3 (LMNSR) addresses this by incorporating natural subregions as a predictor in the linear model. however the same limitations as identified above still apply.

Method 5 (DRC) had extremely weak consistency with the removal rankings of other data cleaning methods, with absolute r-values being less than 0.13. Since the nature of this data cleaning method is based on the formation of clusters based on climatic variables, this method of data cleaning may group data into more broad-scale climatic regions, and more current climatic conditions than those incorporated into the development of natural subregions. This is reflected as based on visual assessment, the spread of clusters (Figure 13) was generally wider than the extent of natural subregions (Figure 2). As a result of the clustering conducted in this method, removal of outliers is also inherently based on larger climatic groupings than natural subregions. The exception to this generalization may be for early blue violet mid bloom and for saskatoon mid bloom since there were more clusters formed through this method, than existing natural subregions. This method also appeared to preferentially remove observations in low density areas (Figure 12). This may be due to the new climatic groupings or clusters, which no longer preferentially retains observations in natural subregions where there are fewer observations, unlike Method 1 (SDiff). Since t-SNE dimensionality reduction is based on an initial random configuration of map points, results may vary each time it is run (van der Maaten and Hinton, 2008). As a result, the change in Moran's I value with each run of Method 5 (DRC) may also vary. An additional consideration for this data cleaning method is the approximately 8 hours of computational time required to run the dimensionality reduction and clustering scripts.

## **5. CONCLUSIONS**

The potential applications for citizen scientist phenological data are diverse and widespread, ranging from climate change to human health. However, prior to any analytical applications, quality assured or quality consistent data is required for reliable data analysis results. A preliminary data screening for "impossible" records or inaccurate data entry may occur at the data entry stage, as it is with the Alberta PlantWatch data. However, additional screening for potentially inaccurate and inconsistent observations is more intensive and depends on a variety of factors, including the capacity for people, time, or finances available through the program administration.

This study evaluated five methods of identifying and removing potentially unreliable data, which may be undertaken by the citizen science program administrator, or the data user. Each method followed a different process based on different practices, or predetermined workflows. While some removed records may be accurate but unusual flowering occurrences resulting from isolated events such as local frost events or variation in microclimate, the goal of data cleaning is to remove potentially inaccurate observations in order to improve data consistency.

The results of different data cleaning methods were not statistically different. All methods demonstrated an increase in spatial autocorrelation, through an improvement (increase) in Moran's I. Two methods that produced the largest increase in Moran's I were Method 1 (SDiff) and Method 5 (DRC). Method 1 (SDiff) preferentially retained records in low density areas, opposite to Method 5 (DRC). Method 1 (SDiff) is also based only on predefined natural subregions that were last updated over 10 years ago using climatic observations from 1961 to 1990 (Natural Regions Committee 2006). Method 5 (DRC) is based on contextual climatic information, specific to the year of the observation, from January 1 up to the date of observation. Method 5 (DRC) also required higher computational time requirements, whereas Method 1 (SDiff) was computationally simple, and does not require additional R software packages.

Where Method 5 (DRC) may be preferred over Method 1 (SDiff) is in identification of potentially similar phenological occurrences, or where data is relatively evenly distributed over a study area. Expanding beyond phenological data cleaning, Method 5 (DRC), as outlined by Mehdipoor et al (2015), could be utilized in situations where contextual factors (in this case climate and location) may drive phenomena. However, the applicability of this method is potentially limited due to larger resource and computational requirements compared to the other data cleaning methods. The use of this method may also be limited by the time requirements of the data cleaner or program administrator.

Method 1 (SDiff) had the largest increase in Moran's I, although only marginally larger than Method 5 (DRC). Method 1 (SDiff) may therefore be preferred over Method 5 (DRC) since it is also an overall simpler method of data cleaning in terms of computational requirements and understanding. For example, citizen scientist coordinators in non-profit organizations may prefer a simpler data evaluation when time and monetary resources are limited. An additional

advantage to Method 1 (SDiff) over Method 5 (DRC) is due to the retention of records in low density areas, for example for use with interpolation applications such as kriging. The final selection of data cleaning method will likely depend on the objectives of data application, the geographic extent of the data required, and the time and resources available to the program administrator or data user.

## 6. LITERATURE CITED

- Alberta Environment (2005). Alberta climate model (ACM) to provide climate estimates (1961-1990) for any location in Alberta from its geographic coordinates. Alberta Environment. <http://www.assembly.ab.ca/lao/library/egovdocs/2005/alene/161251.pdf>. Accessed 7 November 2017.
- Alberta Environment and Parks 2015. Topography. Alberta Environment and Parks. <http://aep.alberta.ca/land/land-industrial/education/physical-land-quality/topography.aspx>. Accessed 4 December 2017.
- Anderson LG, Chapman JK, Escontrela D, Gough LA (2017) The role of conservation volunteers in the detection, monitoring and management of invasive alien lionfish. *Manage Biol Resour* 8:589-598. doi:10.3391/mbi.2017.8.4.14
- Beaubien E, Freeland HJ (2000) Spring phenology trends in Alberta, Canada: links to ocean temperature. *Int J Biometeorol* 44:53-59.
- Beaubien E, Johnson DL (1994) Flowering plant phenology and weather in Alberta, Canada. *Int J Biometeorol* 38:23-27.
- Beaubien E, Hamann A (2011a) Spring flowering response to climate change between 1936 and 2006 in Alberta, Canada. *Biosci* 61:514-524. doi:10.1525/bio.2011.61.7.6
- Beaubien E, Hamann A (2011b) Plant phenology network of citizen scientists: recommendations from two decades of experience in Canada. *Int J Biometeorol* 55:833-841. doi:10.1007/s00484-011-0457-y
- Buldrini F, Simoncelli A, Accordi S, Pezzi G, Dallai D (2015) Ten years of citizen science data collection of wetland plants in an urban protected area. *Bot Lett* 162:365-373. doi:10.1080/12538078.2015.1080187
- Butler D (2009) *asreml: asreml() fits the linear mixed model*. R package version 3.0. [www.vsni.co.uk](http://www.vsni.co.uk)
- Crall AW, Newman GJ, Stohlgren TJ, Holfelder KA, Graham J, Waller DM (2011) Assessing citizen science data quality: an invasive species case study. *Conserv Lett* 4:433-442. doi:10.1111/j.1755-263X.2011.00196.x



- Crall AW, Jarnevich CS, Young NE, Panke BJ, Renz M, Stohlgren TJ (2015) Citizen science contributes to our knowledge of invasive plant distributions. *Biol Invasions* 17:2415-2427. doi: 10.1007/s10530-015-0885-4
- Danielsen F, Jensen PM, Burgess ND et al (2014) A multicountry assessment of tropical resource monitoring by local communities. *Biosci* 64:236-251. <https://doi.org/10.1093/biosci/biu001>
- DataONE (n.d.) DataONE education module: data quality control and assurance. Data Observation network for Earth. [https://www.dataone.org/sites/all/documents/education-modules/pptx/L05\\_DataQualityControlAssurance.pptx](https://www.dataone.org/sites/all/documents/education-modules/pptx/L05_DataQualityControlAssurance.pptx). Accessed 1 November 2017
- Dickinson JL, Zuckerberg B, Bonter DN (2010) Citizen science as an ecological research tool: challenges and benefits. *Annu Rev Ecol* 41:149-172. doi: 10.1146/annurev-ecolsys-102209-144636
- Donaldson J (2012) tsne: t-distributed stochastic neighbor embedding for R (t-SNE). R. Package version 0.1-2. <http://CRAN.R-project.org/package=tsne>
- Donnelly A, Yu R (2017) The rise of phenology with climate change: an evaluation of IJB publications. *Int J Biometeorol* 61 (Suppl 1):S29-S50. doi: 10.1007/s00484-017-1371-8
- Fore LS, Paulsen K, O’Laughlin K (2001) Assessing the performance of volunteers in monitoring streams. *Freshw Biol* 46:109-123. doi:10.1046/j.1365-2427.2001.00640.x
- Fraley C, Raftery AE, Murphy B, Scrucca L (2012) mclust version 4 for R: normal mixture modeling for model-based clustering, classification, and density estimation technical report No. 597. Department of Statistics, University of Washington
- Fuccillo KK, Crimmins TM, de Riviera CE, Elder TS (2014) Assessing accuracy in science-based plant phenology monitoring. *Int J Biometeorol* 59: 917-926. doi: 10.1007/s00484-014-0892-7
- Gallo T, Waitt D (2011) Creating a successful citizen science model to detect and report invasive species. *Biosci* 61:459-465. doi:10.1525/bio.2011.61.6.8
- Gardiner MM, Allee LL, Brown PMJ, Losey JE, Roy HE, Smyth RR (2012) Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. *Front Ecol Environ* 10:471-476. doi:10.1890/110185
- Granek EF, Madin EMP, Brown MA et al (2008) Engaging recreational fishers in management and conservation: global case studies. *Conserv Biol* 22: 1125-1134. doi: 10.1111/j.1523-1739-2008.00977.x
- Havens K, Vitt P, Masi S (2012) Citizen science on a local scale: the Plants of Concern program. *Front Ecol Environ* 10:321-323. doi: 10.1890/110258
- Hufkens K (2017) khufkens/daymetr: download daymet data using R. Zenodo. <http://doi.org/10.5281/zenodo.437886>.
- Hugo S, Altwegg R (2017) The second Southern African bird atlas project: causes and consequences of geographical sampling bias. *Ecol and Evol* 7:6839-6849. doi: 10.1002/ece3.3228

- Ingwell LL, Preisser EL (2011) Using citizen science programs to identify host resistance in pest-invaded forests. *Conserv Biol* 25:182-188. doi: 10.1111/j.1523-1739.2010.01567.x
- Intergovernmental Panel on Climate Change (2007) *Climate Change 2007: Synthesis Report. Contribution of Working Groups I, II, and III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Core Writing Team, Pachauri RK, Reisinger A (eds). IPCC, Geneva, Switzerland, 104 pp.
- Javari M (2017) Assessment of temperature and elevation controls on spatial variability of rainfall in Iran. *Atmos* 8. doi:10.3390/atmos8030045
- Jiguet F, Devictor V, Julliard R, Couvet D (2012) French citizens monitoring ordinary birds provide tools for conservation and ecological sciences. *Acta Oecol* 44:58-66. <https://doi.org/10.1016/j.actao.2011.05.003>
- Kelling S, Lagoze C, Wong W, Yu J, Damoulas T, Gerbracht J, Fink D, Gomes C (2013) eBird: a human/computer learning network to improve biodiversity conservation and research. *AI Mag* 34. <https://doi.org/10.1609/aimag.v34i1.2431>
- Kosmala M, Wiggins A, Swanson A, Simmons B (2016) Assessing data quality in citizen science. *Front Ecol Environ* 14: 551-560. doi: 10.1002/fee.1436
- La Sorte FA, Thompson III FR (2007) Poleward shifts in winter ranges of North American birds. *Ecol* 88:1803-1812. doi:10.1890/06-1072.1
- Latta G, Temesgen H, Barrett TM (2009) Mapping and imputing potential productivity of Pacific Northwest forests using climatic variables. *Can J For Res* 39:1197-1207. <https://doi.org/10.1139/X09-046>
- Levrel H, Fontaine B, Henry PY, Jiguet F, Julliard F, Kerbiriou C, Couvet D (2010) Balancing state and volunteer investment in biodiversity monitoring for the implementation of CBD indicators: a French example. *Ecol Econ* 69:1580-1586. doi: 10.1016/j.ecolecon.2010.03.001
- Lewis DE (2003) *Presidents and the politics of agency design: political insultation in the United States government bureaucracy, 1946-1997*. Stanford, California
- MacKenzie CM, Murray G, Primack R, Weihrauch D (2017) Lessons from citizen science: assessing volunteer-collected plant phenology data with Mountain watch. *Biol Conserv* 208:121-126. doi: 10.1016/j.biocon.2016.07.027
- Marchante H, Morais MC, Gamela A, Marchante E (2017) Using a webmapping platform to engage volunteers to collect data on invasive plants distribution. *Trans GIS* 21:283-252. doi: 10.1111/tgis.12198
- Mehdipoor H, Zurita-Milla R, Rosemartin A, Gerst KL, Weltzin JF (2015) Developing a workflow to identify inconsistencies in volunteered geographic information: a phenological case study. *Plos One* 10: doi: 10.1371/journal.pone.0140811
- Mengersen K, Peterson EE, Clifford S, Ye N, Kim J, Bednarz T, Brown R, James A, Vercelloni J, Pearse AR, Davis J, Hunter V (2017) Modelling imperfect presence data obtained by citizen science. *Environmetrics* 28. <https://doi.org/10.1002/env.2446>

- Menzel A (2003) Europe. In: Schwartz MD (ed) *Phenology: An Integrative Environmental Science. Tasks for Vegetation Science*, vol 39. Springer, Dordrecht, pp 45-56
- McKinley DC, Miller-Rushing AJ, Ballard HL et al (2017) Citizen science can improve conservation science, natural resource management, and environmental protection. *Biol Conserv* 208: 15-28. <https://doi.org/10.1016/j.biocon.2016.05.015>
- Miller-Rushing A, Primack R, Bonney R (2012) The history of public participation in ecological research. *Front Ecol Environ* 10:285-290. doi: 10.1890/1102798
- Morii Y, Nakano T (2017) Citizen science reveals the present range and a potential native predator of the invasive slug *Limax maximus* Linnaeus, 1758 in Hokkaido, Japan. *Bioinvasions Rec* 6: 181-186. doi: 10.3391/bir.2017.6.3.01
- Natural Regions Committee (2006). *Natural Regions and Subregions of Alberta*. Compiled by D.J. Downing and W.W. Pettapiece. Edmonton. Pub. No. T/852. Alberta Environment, Government of Alberta, Edmonton, AB
- Nature Canada (2010) *Plantwatch Canada in bloom!* Environment Canada. Catalogue No.: En4-111/2009E, ISBN 978-1-100-12387-5
- Paradis E, Claude J, Strimmer K (2004) APE: analysis of phylogenetics and evolution in R language. *Bioinformatics* 20:289-290.
- Ranjitkar S (2013) Effect of elevation and latitude on spring phenology of rhododendron and Kanchenjunga conservation area, East Nepal. *Int J Appl Sci and Biotech* 1: 253-257. Doi: 10.3126/ijasbt.v1i4.9154
- Rathcke B, Lacey EP (1985) Phenological patterns of terrestrial plants. *Annual Rev Ecol Syst* 16:179-214.
- Reichhardt T (1994) US ecological survey runs political gauntlet. *Nature* 367:400-400.
- Rosemartin AH, Crimmins TM, Enquist CAF et al (2014) Organizing phenological data resources to inform natural resource conservation. *Biol Conserv* 173:90-97. doi:10.1016/j.biocon.2013.07.003
- R Development Core Team (2014) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- Savolainen O, Pyhäjärvi T, Knürr T (2007) Gene flow and local adaptation in trees. *Annu Rev Evol Syst* 38:595-619. <https://doi.org/10.1146/annurev.ecolsys.38.091206.095646>
- Schwartz MD, Hanes JM, Liang L (2014) Separating temperature from other factors in phenological measurements. *Int J Biomet* 58: 1699-1704. <https://doi.org/10.1007/s00484-013-0723-2>
- Shirk JL, Ballard HL, Wilderman CC et al (2012) Public participation in scientific research: a framework for deliberate design. *Ecol and Soc* 17:29. <http://dx.doi.org/10.5751/ES-04705-170229>
- Silvertown J, Buesching CD, Jacobson SK, Rebelo T (2013) Citizen science and nature conservation. In: Macdonald DW and Willis KJ (ed) *Key Topics in Conservation Biology* 2, 1<sup>st</sup> edn. Wiley, New York, pp 127-142

- Smith P, Spalding DAE, Davidson RB, Harrison RO (2017) Alberta, Canada. Encyclopedia Britannica. <https://www.britannica.com/place/Alberta-province>. Accessed 4 December 2017.
- Sparks TH, Huber K, Tryjanowski P (2008) Something for the weekend? Examining the bias in avian phenological recording. *Int J Biomet* 52:505-510. doi: 10.1007/s00484-008-0146-7
- Strong WL, Leggat KR (1992) Ecoregions of Alberta. Alberta Forestry, Lands and Wildlife. Pub No. T/245. ISBN. 0-86499-840-6
- Thornton PE, Thornton MM, Mayer BW, Wilhelmi N, Wei Y, Devarakonda R, Cook RB (2016) Daymet: daily surface weather data on a 1-km grid for North America, Version 3 ORNL DAAC, Oak Ridge, Tennessee, USA. Accessed June 5, 2017. Time period: 1987-01-01 to 2016-12-31. Spatial range: N=59.82, S=49.13, E=-109.22, W=-119.67. <http://dx.doi.org/10.3334/ORNLDAAC/1219>
- Travel Alberta (2017) Weather and Climate. Travel Alberta. <https://www.travelalberta.com/ca/plan-your-trip/weather-climate/>. Accessed 4 December 2017
- USA National Phenology Network (n.d.) How to observe. USA National Phenology Network. from <https://www.usanpn.org/nn/guidelines>. Accessed 2 November 2017.
- van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Machin Learn Res* 9:1-48.
- Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Econ Geog* 46: 234-240. doi:10.2307/143141
- Vander Stelt E, Fant JB, Masi S, Larkin DJ (2017) Assessing habitat requirements and genetic status of a rare ephemeral wetland plant species, *Isoëtes butleri* Engelm. *Aquat Bot* 138:74-81. <https://doi.org/10.1016/j.aquabot.2017.01.002>
- Wagner FH (1999) Whatever happened to the National Biological Survey? *Biosci* 49:219-222. <https://doi.org/10.2307/1313512>
- Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer-verlag New York
- Zhou L, Braun WJ, Woolford DG, Wotton M (2009) A simulation study of predicting flush date. *Commun in Stat – Simul and Comput* 38: 1071-1082. doi: 10.1080/03610910902785738
- Zhou X, Lin H (2008) Moran's I. In: Encyclopedia of GIS. Springer-Verlag. [https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-35973-1\\_817](https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-35973-1_817). Accessed 13 Sept 2017

## Appendix A – Data trends before and after cleaning by species and phase

**Table 7.** Summary statistics before and after data cleaning for citizen scientist collected phenological data in Alberta, Canada 1987-2016.

Species	Phase	Statistic	All Records	Method 1 (Standardized difference)	Method 2 (Linear model)	Method 3 (Linear model by natural subregion)	Method 4 (Observer correlation)	Method 5 (Dimensionality reduction and clustering)
Aspen	First Bloom	Range	63-197	63-150	68-141	68-141	63-150	63-197
		No. records	1400	1330	1330	1330	1330	1330
		Mean	106.3	106.1	106.2	106	106.3	106.2
		Median	107	106	106.5	106	106	106
		Stand. Dev.	12.7	11.7	11.1	11.1	12.5	12.8
	Mid Bloom	Range	66-155	66-155	72-143	72-143	66-155	66-155
		No. records	1183	1124	1124	1124	1124	1124
		Mean	110.5	110.2	110.2	110.2	110.4	110.2
		Median	110	110	110	110	110	110
		Stand. Dev.	12	11.5	10.8	10.8	12.1	12.1
Early blue violet	First Bloom	Range	89-183	89-183	99-160	99-166	89-183	89-183
		No. records	1754	1667	1667	1667	1667	1667
		Mean	130.8	130.9	130.7	130.8	131.1	130.6
		Median	130	130	130	130	130	130
		Stand. Dev.	10.7	10.3	9.21	9.26	10.6	10.7
	Mid Bloom	Range	95-188	100-188	107-186	106-171	95-188	95-188
		No. records	1539	1462	1462	1462	1462	1462
		Mean	136.3	136.2	136	135.9	136.5	136.3
		Median	135	135	135	135	136	136
		Stand. Dev.	11	10.4	9.56	9.4	11	10.9
Saskatoon	First Bloom	Range	95-190	108-190	108-174	108-174	95-190	108-178
		No. records	2200	2090	2090	2090	2090	2090
		Mean	137.8	137.8	137.7	137.7	137.8	138
		Median	138	138	137	137	138	138
		Stand. Dev.	9.8	9.28	8.71	8.79	9.68	9.17

	Mid	Range	93-179	110-179	119-175	119-175	93-179	93-176
	Bloom	No. records	1946	1849	1849	1849	1849	1849
		Mean	141.3	141.3	141	141.1	141.4	141
		Median	141	141	141	141	141	141
		Stand. Dev.	9.91	9.48	8.95	8.99	9.83	9.65
Chokecherry	First	Range	101-179	101-179	114-173	114-173	101-179	101-175
	Bloom	No. records	1495	1421	1421	1421	1421	1421
		Mean	146	146.3	146.3	146.3	146.1	146.2
		Median	146	147	146	147	147	146
		Stand. Dev.	10.1	9.62	9.17	9.21	9.95	9.92
	Mid	Range	106-192	118-187	123-182	123-182	106-192	106-190
	Bloom	No. records	1285	1220	1220	1220	1220	1220
		Mean	150.3	150.6	150.5	150.5	150.6	150.4
		Median	150	150	150	150	151	150
		Stand. Dev.	10.2	9.59	9.05	9.18	10.2	10.1

**Appendix B** – Moran’s I for each cleaning method by species and phase

**Table 8.** Moran’s I statistic before and after data cleaning for citizen scientist collected phenological data in Alberta, Canada 1987-2016.

Species	Phase	Statistic	All Records	Method 1 (Standardized difference)	Method 2 (Linear model)	Method 3 (Linear model by natural subregion)	Method 4 (Observer correlation)	Method 5 (Dimensionality reduction and clustering)	
Aspen	First Bloom	Range	-0.0739-0.165	-0.0739-0.165	-0.208-0.159	-0.0933-0.142	-0.0801-0.172	-0.114-0.280	
		No. records	30	30	30	30	30	30	
		Mean	0.0248	0.0248	0.0207	0.0214	0.0233	0.0254	
		Median	0.0131	0.0131	0.0255	0.0174	0.0256	0.0207	
		Stand. Dev.	0.0648	0.0648	0.0757	0.0602	0.0649	0.0802	
	Mid Bloom	Range	-0.118-0.164	-0.245-0.187	-0.118-0.156	-0.118-0.137	-0.118-0.155	-0.121-0.213	
		No. records	30	30	30	30	30	30	
		Mean	0.0224	0.0303	0.0384	0.0286	0.0217	0.0323	
		Median	0.0226	0.0204	0.0365	0.0255	0.0292	0.0288	
		Stand. Dev.	0.0661	0.0865	0.0624	0.0605	0.0657	0.0732	
	Early blue violet	First Bloom	Range	-0.0446-0.207	-0.0774-0.220	-0.107-0.173	-0.107-0.176	-0.0725-0.185	-0.0504-0.210
			No. records	30	30	30	30	30	30
Mean			0.0322	0.0385	0.02	0.0219	0.0353	0.0342	
Median			0.0196	0.0311	0.0208	0.0291	0.0339	0.024	
Stand. Dev.			0.058	0.065	0.0658	0.0656	0.0611	0.0613	
Mid Bloom		Range	-0.0774-0.223	-0.0774-0.249	-0.179-0.213	-0.179-0.0937	-0.0774-0.211	-0.0774-0.235	
		No. records	30	30	30	30	30	30	
		Mean	0.0255	0.0363	0.0188	0.00576	0.0303	0.0312	
		Median	0.0196	0.0317	0.0188	0.0171	0.0226	0.0282	
		Stand. Dev.	0.0554	0.0624	0.0676	0.0532	0.0564	0.0641	
Saskatoon	First Bloom	Range	-0.0795-0.236	-0.0942-0.255	-0.0795-0.359	-0.0795-0.359	-0.0513-0.244	-0.00280-0.246	
		No. records	30	30	30	30	30	30	
		Mean	0.0642	0.0787	0.0917	0.0856	0.0734	0.0818	
		Median	0.0531	0.0729	0.0801	0.0589	0.0615	0.0616	
		Stand. Dev.	0.0687	0.0754	0.0874	0.087	0.0665	0.0676	

	Mid	Range	-0.0871-0.207	-0.169-0.207	-0.0309-0.277	-0.0280-0.230	-0.141-0.252	-0.087-0.207
	Bloom	No. records	30	30	30	30	30	30
		Mean	0.0542	0.0603	0.076	0.0734	0.0605	0.0603
		Median	0.0501	0.0634	0.06	0.0538	0.0562	0.0531
		Stand. Dev.	0.0714	0.0774	0.0703	0.064	0.0798	0.0695
Chokecherry	First	Range	-0.0714-0.218	-0.0714-0.328	-0.0714-0.178	-0.0634-0.280	-0.0830-0.218	-0.123-0.218
	Bloom	No. records	30	30	30	30	30	30
		Mean	0.0361	0.0443	0.0299	0.0378	0.0309	0.0389
		Median	0.0203	0.0221	0.0149	0.0146	0.0175	0.026
		Stand. Dev.	0.0686	0.0786	0.0662	0.0756	0.0737	0.0805
	Mid	Range	-0.105-0.258	-0.146-0.258	-0.115-0.258	-0.0924-0.258	-0.117-0.277	-0.0846-0.258
	Bloom	No. records	30	30	30	30	30	30
		Mean	0.0248	0.0282	0.0332	0.0389	0.0218	0.0336
		Median	0.00485	0.00751	0.0196	0.0255	0.00496	0.0205
		Stand. Dev.	0.07659	0.0826	0.0824	0.0803	0.0792	0.0804