

MINING SPATIAL AND SPATIO- TEMPORAL CO-LOCATIONS IN VOLUNTEERED PHENOLOGICAL OBSERVATIONS

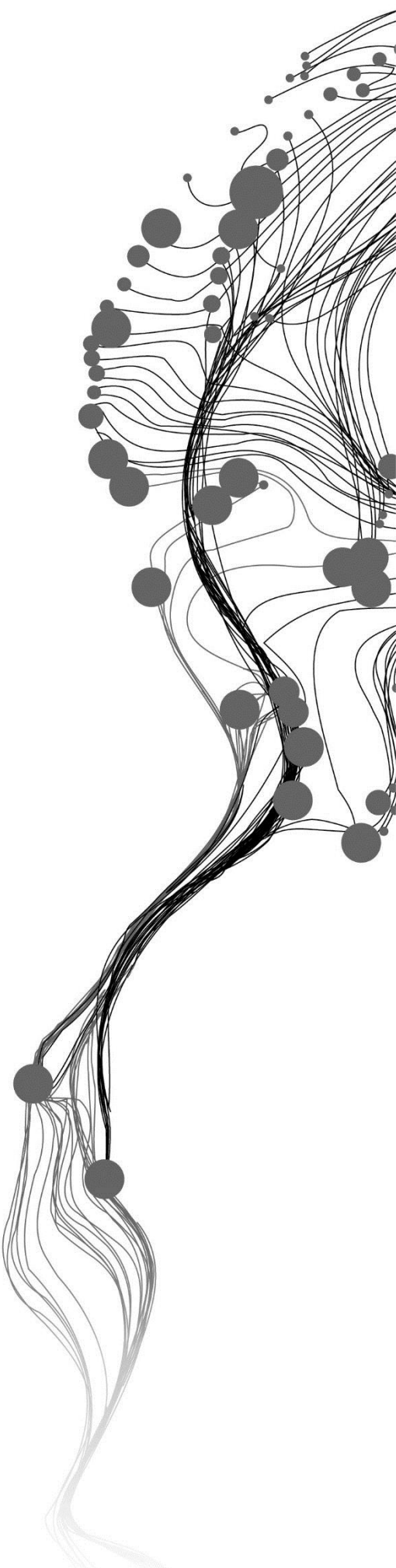
NA JIN

February, 2016

SUPERVISORS:

Dr. R. Zurita-Milla

Ms ir. P.W.M. Augustijn



MINING SPATIAL AND SPATIO- TEMPORAL CO-LOCATIONS IN VOLUNTEERED PHENOLOGICAL OBSERVATIONS

NA JIN

Enschede, The Netherlands, February, 2016

Thesis submitted to the Faculty of Geo-Information Science and Earth
Observation of the University of Twente in partial fulfilment of the requirements
for the degree of Master of Science in Geo-information Science and Earth
Observation.

Specialization: Geo-informatics

SUPERVISORS:

Dr. R. Zurita-Milla

Ms ir. P.W.M. Augustijn

THESIS ASSESSMENT BOARD:

Dr.ir. R.A. de By (Chair)

Dr.ir. A.J.H. van Vliet (External Examiner, Wageningen University)

DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

ABSTRACT

Spatio-temporal analysis is regarded as the core of GIS because it is where the spatial and temporal patterns, relationships, geographic process can be found to obtain the insights in how the world is evolving. Data mining techniques used in spatio-temporal analysis emerges as the basis for decision-making across a diverse range of domains including phenology. One of the common source of phenological data is Volunteered Geographical Information which has been used to study the impact of climate change. Based on volunteered phenological data, many analysis can be made to study the impact of climate change on different species. In the past, many researches focused on phenological event of one species ignoring the interrelations among species. However, the relationships among species are important, because the underlying processes and interactions among species indicates their response to the environment.

The main objective of this study is to discover and map spatial and spatio-temporal co-locations using volunteered phenological observations. This means finding the genus that tend to occur next to each other and finding the spatially co-located genus that have the same timing for a certain phenophase. In this research, we applied Apriori algorithm to find the co-locations in USA from volunteered phenological observations. Because the co-location rules cannot be derived directly from the phenological observations directly. Two different ways of neighbourhood generation to mine spatial co-location were used and discussed. Firstly, the neighbourhood was defined using distance as the radius. Secondly, the neighbourhood was defined both on distance and elevation difference among different observations. To mine the spatio-temporal co-locations, the neighbourhood was defined by distance and the temporal element was defined by the overlapped range between two genus for the phenophase. The way to define appropriate neighbourhood size and configure the best minimum support and confidence to find the meaningful co-location rules were illustrated in my study. Besides, we discussed the method of presenting the spatial and spatio-temporal co-location relationship in graph-based techniques and in maps.

From the mined co-location result, we found that genus Forsythia, Cornus and Syringa tend to occur nearby when using distance as the radius for neighbourhood. Syringa and Forsythia, Syringa and Cornus, Syringa and Cercis, these genus pairs tend to co-locate within a radius neighbourhood and within the same elevation difference. Moreover, the suggestion of aggregation the duration of phenophase for species to mine the spatio-temporal co-locations was carried out. Further study of this topic would enable to predict the possible area or the phenophase for a certain genus for which missing records exist in the database based on the mined spatial and spatio-temporal co-location rules.

Key words: Co-location mining, Volunteered Phenological observations, Spatio-temporal analysis

ACKNOWLEDGEMENTS

First and foremost, I would like to express my greatest gratitude to my first supervisor, Dr.R.Zurita-Milla for his constant support of my MSc research. His guidance, patience, enthusiasm and immense knowledge helped me to keep doing the research on the right track and manage to finalizing the thesis eventually. It is the way. Frequent enlightening discussion with him encouraged me to make constant progress on my research. Besides, I really appreciated my second supervisor Ir. P.W.M. Augustijn (Ellen-Wien) for her supervision all the time of research and writing of my thesis, and the valuable remarks and encouragement.

Furthermore, I would like to thank the rest of my thesis committee: Prof.dr. M.J. Kraak (Menno-Jan), Dr.ir. R.A. de By and Drs.J.P.G. Bakx (Wan) for their hard questions and insightful comments during the proposal and mid-term defence.

And I would like to acknowledge the USA-NPN website for the data support and replying my e-mail in time to help me solve the problems of the data.

My sincere thanks also goes to ITC, especially the department of GFM. Thanks to this program, I have spent meaningful one year and a half accumulating and learning new knowledge and skills, which is helpful for my research and also for the further work in the future.

Last but not least, I would like to thank my parents, my friends for theirs support, love and unconditional care whenever I felt stressful or upset.

TABLE OF CONTENTS

List of figures	iv
List of tables	v
1. Introduction.....	1
1.1. Motivation and Problem Statement	1
1.2. Research Identification	2
1.3. Innovation Aimed at.....	3
2. Literature Review.....	4
2.1. Review of Spatial Co-location Mining Method.....	4
2.2. Volunteered Phenological Observations.....	8
2.3. Spatio-temporal Analysis in Volunteered Geographic Information	8
3. Materials and Method	10
3.1. Materials	10
3.2. Method	11
4. Result and Discussion.....	17
4.1. Data Preparation.....	17
4.2. Spatial Co-location Mining	18
4.3. Spatio-temporal Co-location Mining	30
5. Conclusions and Recommendations	37
5.1. Conclusions	37
5.2. Recommendations.....	39
List of references	41
Appendix.....	44

LIST OF FIGURES

Figure 2.1 Transaction data sample.....	5
Figure 2.2 Apriori algorithm.....	7
Figure 3.1 Method workflow to answer the research questions	11
Figure 4.1 Histogram of the observations available for each year	17
Figure 4.2 Total number of unique locations for each year	17
Figure 4.3 Total number of genus types and species types for each year	18
Figure 4.4 Process from transaction data to spatial transaction data (Sample dataset).....	19
Figure 4.5 Spatial neighbourhood transaction data based on distance and elevation	20
Figure 4.6 Scatter plot of spatial co-location rules for different size of neighbourhood (Distance)	21
Figure 4.7 Graph-based spatial co-location based on distance rules visualization using items as vertices for different neighborhood size.....	23
Figure 4.8 Spatial co-location rules map generated by distance when the distance is 20km.....	24
Figure 4.9 Scatter plot of spatial co-location rules for different size of neighborhood (Distance and elevation)	25
Figure 4.10 Graph-based visualization using items as vertices for spatial co-location rules from distance and elevation	27
Figure 4.11 Graph-based visualization using item sets as vertices for spatial co-location rules from distance and elevation	28
Figure 4.12 Spatial co-location rules map generated by distance and elevation when the distance is 20km	29
Figure 4.13 Coordinates conversion in spatio-temporal co-location mining table.....	30
Figure 4.14 Spatio-temporal transaction data(sample)	30
Figure 4.15 Scatter plot of spatial and temporal co-location rules for different size of neighborhood.....	32
Figure 4.16 Graph-based visualization using items as vertices for spatial co-location rules from distance and elevation	33
Figure 4.17 Graph-based visualization using itemsets as vertices for spatial co-location rules from distance and elevation	34
Figure 4.18 Spatio-temporal co-location rules maps when the neighborhood size was 20km.....	36
Figure 4.19 Genus pairs phenophase time range for spatial-temporal co-location rules	36

LIST OF TABLES

Table 3.1 Records showing instances in volunteer phenological observation data from USA-NPN 10

Table 4.1 Median support derived from support histogram shown in Table (a) and the best minimum
support for the spatial co-location rules based on distance shown in Table (b) 22

Table 4.2 Median support derived from support histogram shown in Table (a) and the best minimum
support for the spatial co-location rules based on distance and elevation shown in Table (b) 26

Table 4.3 Median support derived from support histogram shown in Table (a) and the best minimum
support for the spatial co-location rules based on distance shown in Table (b) 31

1. INTRODUCTION

1.1. Motivation and Problem Statement

Spatio-temporal analysis is often referred to as the core of GIS because it is where the spatial and temporal patterns, relationships and processes of geographic, cultural and biological phenomena are explored at scientific level to acquire insights into how the world is evolving and changing (Longley & Batty, 2003). With the advance of information communication and the improvement in mobile location-aware techniques, volunteered geographic information has become a novel source of data providing scientists with massive amounts of data (Mehdipoor et al., 2015). The speed of vast and voluminous spatio-temporal data generation is faster than their analysis, so what is lacking in GIS is analytical tools for discovering new insights and information (Estivill-castro & Murray, 1994). However, hidden and unexpected information in large spatial and temporal databases cannot be discovered easily by traditional statistical methods for acquiring a prior hypothesis or strict assumptions (Miller & Han, 2009a). In that case, spatio-temporal data mining techniques are necessary to reveal insights and potential useful knowledge (Hagenauer & Helbich, 2013).

The application of data mining techniques in spatio-temporal analysis emerges as the basis for decision-making across a diverse range of domains (Longley & Batty, 2003). For example, association rule mining in spatial and temporal database was applied to find the relationship among socioeconomic characteristics and land cover change in Colorado, USA (Mennis & Liu, 2005). Movement patterns in space and time, like yearly migration patterns or daily commuting patterns, were extracted and discovered by spatio-temporal association rule mining method to study human behaviour or to support traffic management (Gudmundsson, Laube, & Wolle, 2008). It offered the impetus for the methods to analyse massive amounts of spatio-temporal data, so the data mining methods could help to the specificities of phenology. Phenology is defined as the study of plant and animal life events in recurring periods. For instance, the analysis on synoptic spatiotemporal phenological patterns over large area using self-organizing map (SOM) and Sammon's projection method that involves unsupervised data mining (Zurita-Milla, Van Gijssel, Hamm, Augustijn, & Vrieling, 2013); understanding topographic control on climate-induced inter-annual vegetation variability over the United States by regression tree induction which is a data mining method categorized under classification and prediction (White, Kumar, & Tcheng, 2005). Many factors, such as climate variation and its modification of abundance, diversity and interaction of species have impact on the change of the phenological events (Betancourt, Schwartz, & Breshears, 2007). The significance of phenology in terms of global change has been realized widely (Richardson et al., 2013). The changes in the timing of phenophase of species like flowering for plants, or earlier breeding for birds, reflect the impact of climate changes (Walther et al., 2002). For the high sensitivity of phenology to climate change, it can be applied in many areas, not only land management, but also human health as well as numerous ecosystem services (Schröter et al., 2005). More specifically, land surface phenology is used to analyse agricultural land cover change in space and time (De Beurs & Henebry, 2004), and the impact of pollen on allergic disorders can be studied by observing phenological events (Huynen, Menne, Behrendt, & Bertollini, 2003). In ecology domain, the study of phenological shifts helps to track the changing statements of ecosystem and support a better understanding of biological underpinning (Mooney et al., 2009).

The early concerns and researches in phenological spatio-temporal analysis mainly focused on the actual day of each phenological event. However, the understanding of the interrelations between different phenophases of species plays a key role in monitoring climate change as well. Because phenology will affect

the competitive interactions among species and across trophic levels, the position it occupies in a food chain, leading to different distribution of species (Chuine, 2010). More specifically, there are possibly animals occurring in the neighbourhood of certain plant species forming regular patterns. Hence, the relationship of the interesting and closely related distribution among different species will reveal the impact of climate change.

Instead of concentrating on phenological events of only one plant species or only one animal species, the reason for discovering the co-location occurrences between different species is that some specific species are sensitive to climate change and have phenophases clearly responding to climatic variation, but some species' phenophases are not clearly related to climate change (Möller & Gläßer, 2011). Therefore, the study of relationships between species give much more support to detect, monitor and forecast ecological change. From the perspective of ecologists, the exploration of spatial co-location occurrences pattern indicates underlying processes and interactions among various species and their movement or response to the environment. It implies for both theoretical problems, like environmental heterogeneity or resource availability and application issues as threaten species management (Perry & Dixon, 2002). For example, one certain plant may always occurs in the region where some animals assemble and some diseases tend to break out in the place where some insects dominate very often (Wang, Zhou, Lu, & Yip, 2009)

It is important to understand the connection of environmental features with phenological events among different species. Phenological observations are regarded as a worthwhile source of interactions on the ground surface (Schwartz, 1994). Using space-borne and airborne remote sensing devices, phenological data can be obtained, of which quality is influenced by many factors, like atmospheric disturbances, cloud cover (Studer et al., 2007). With the development of Web 2.0 technology and the prosperity of mobile devices, obtaining data based on Volunteer Geographic Information (VGI) has become a popular source of spatio-temporal dataset (Batarseh, 2014). Compared with traditional approaches, VGI-based data not only allows broader spatial extent and higher temporal frequency, but also involves the human participation, and is time-saving and less costly. Volunteered phenological observations have been used to detect and predict spatio-temporal environmental changes under the impact of land use and climate change (New, 2005). However, there is no guarantee that the phenological information for an existing species in the same place will be collected each year. Both the completeness and the continuity of observations play important roles in understanding the roles of climatic variability among species interactions. The incomplete nature of existing phenological database offers limited opportunities for the analysis across temporal dimension (Mayer, 2010). The lack of data in important locations will lead to spatially-biased monitoring results (Goodchild, 2007).

In this context, the study is motivated towards mining co-location patterns in space and time among different species based on volunteered phenological observations.

1.2. Research Identification

1.2.1. Research Objectives

The discovery of interesting relationship between species that tend to occur in the neighbourhood from each other, and the relationship among their phenological phases of species will provide insights on interactive processes among species to assistant better understanding of the impact of climate change. Therefore, the main objective of this research is to mine spatial and spatio-temporal co-locations based on volunteered phenological observations.

- Mining and mapping spatial co-location patterns based on the locations of the phenological observations collected by volunteers.
- Mining and mapping spatio-temporal co-location occurrences patterns based on the timing of the phenophases reported by volunteers.

1.2.2. Research Questions

1. In spatial dimension

- 1.1. How to prepare the transaction data in space?
- 1.2. How to define the neighbourhood in space?
- 1.3. How to define the size of the neighbourhood of species to fix the co-location mining area?
- 1.4. Which species tend to appear near to each other?
- 1.5. How to visualize the spatial co-locations?
- 1.6. How to define the meaningful spatial co-location rules?

2. In spatio-temporal dimension

- 2.1. How to prepare the transaction data in both space and time?
- 2.2. How to define the neighbourhood in spatio-temporal way?
- 2.3. Which of the spatially co-located species have similar timing of phenophases?
- 2.4. Which phenophase of the spatially co-located species have when they co-occur with each other?
- 2.5. How to map the spatio-temporal co-locations?
- 2.6. How to define the meaningful spatial co-location rules?

1.3. Innovation Aimed at

From a geoinformatics point of view, this study proposes the use of an existing database for innovative application. In this case, the database was collected by volunteers to study the impact of climate change and not directly to study species distribution.

From a phenological point of view, this research will emphasize the colocation pattern among different species of different timing of phenophase, aiming at finding the relationship among different individuals unlike previous researches that focus on the timing of phenological events (for group of species).

2. LITERATURE REVIEW

2.1. Review of Spatial Co-location Mining Method

Spatial co-location rules represents relationships among the events that occur next to each other (Shekhar & Huang, 2001). Many studies describing methods to mine spatial co-location patterns from spatial data sets can be found in literature. These methods can be divided into spatial statistics and data mining approaches (Huang, Shekhar, & Xiong, 2004). In spatial statistics approaches, measures of spatial correlation are applied to characterize the relationship between different spatial features (Shekhar & Huang, 2001). Spatial correlation is quantified by chi-square tests, correlation coefficients and regression models. However, the computation of spatial correlation measures for all co-location patterns is computationally costly when it comes to a huge number of spatial candidates (Miller & Han, 2009).

Data mining approaches can be categorized into clustering-based map overlay and association rule mining approaches (Shekhar & Huang, 2001). In clustering-based method, it regards each attribute as a map layer and at the same time consider spatial clusters as candidates from different layers to mine the association. This approach was applied in the study of finding potential rules for crime occurrences. Estivill-castro & Murray (1994) used this method to discover a relationship between the location of stolen cars and subway stations, considering each attribute like crime occurrences, distribution of stolen cars, subway lines and parks as map layers that are clustered spatially. However, the disadvantages of this method are as follows: the instances of each spatial features are considered to be clustered, but in reality, species observations are randomly sampled in space; its high sensitivity to the choice of clustering algorithm and presence of noise makes it hard to control the accuracy of results (Xiong et al., 2004).

In association rule mining, this method is applied to discover frequent patterns and associations in the transaction databases where sets of items are stored (Kotsiantis & Kanellopoulos, 2006). it is one of an important data mining technique, which is used to find out the rules that meet user-defined minimum support and confidence (Kotsiantis & Kanellopoulos, 2006). Mined association rules are in the form of $A \rightarrow B$, where A and B refers to sets of predicates (Mennis & Liu, 2005), which can be goods stored in a supermarket. For instance, when the transaction database is built based on a supermarket, the association rule would be as “the customer that buy bread tends to buy butter as well”. To find out the association rules, the process can be split into two steps: the first step is to find the frequent items of which occurrences meet the predefined support; the second one is to generate association rules from frequent items based on the minimum user-defined confidence (Kotsiantis & Kanellopoulos, 2006). By association rule-based approach, it has helped retailers who interested in finding items frequently bought together to make store arrangements and promote products together (Shekhar & Huang, 2001). Originally, non-spatial association rule mining is generated rules from categorical data, having nothing to do with numeric data such as distance among geographic coordinates (Mennis & Liu, 2005). Spatial databases usually contain traditional data and geographic information about the corresponding data as well (Shekhar & Huang, 2001). As an extension, spatial association rules indicate object/predicate relationships that contain spatial predicate (Koperski & Han, 1995). To derive association rules in spatial database, the discretization of numeric data into ordinal categories can be implemented and the rules are mined from the data categorized (Agrawal & Srikant, 1994). For example, the distance usually is used

to defined the radius of an neighbourhood, so the discretization can be made based on distance interval, like within 5km, 10km and etc. In that way, in a spatial database all the spatial features can be regarded as goods sold in the supermarket. Similarly, the spatial and temporal co-locations refer to the relationship among the events that happen in a spatial neighbourhood and in a temporal neighbourhood as well (Verhein & Chawla, 2006). An example for the association rules mined in temporal database, instead of finding the association between bread and butter only, the time interval for buying this two products is also mined from the rule (Mennis & Liu, 2005).

One of association rule-based approach is transaction-based method. In market basket analysis, the transactions mean sets of item types that are bought together, which only concentrates on the frequent items (Shekhar & Huang, 2001). Because of the definition of transactions over space, an Apriori-like algorithm can be used (Agrawal & Srikant, 1994). The transactions can be defined by reference feature centric model or data-partition approach. The data-partition approach tends to measure colocations by grouping the instances into disjoint partitions which is useful when it comes to identifying the regions that maximize the co-location (Morimoto, 2001). Although transactions are independent because they are not sharing instances of items and not sharing spatial relationships (e.g. neighbour) with each other (Shekhar & Huang, 2001), the transaction can be rearranged based on the spatial relationship like distance to obtain new spatial transactions.

Generally, in transaction-based method, each transaction contains many items. And each transaction is stored as one row in the database. In Figure 2.1, each letter presents an item and stored in each transaction. The frequent itemsets referred to the combination of the items, like {T,B}, {T ,B}, that occurring multiple times in the transaction database.

TID	Items
1	TCB
2	TCQ
3	CB
4	TB
5	CBQ

Figure 2.1 Transaction data sample

Support is the proportion of transaction in the transaction database. For example, in Figure 2. 1 Support (TC) is 0.4, because the support count of itemsets TC's co-occurrence is 2, and the total count of transaction is 5. It means the possibility of TC's appear together is 40%.

If $X \rightarrow Y$ is the mined rule, confidence is defined as the proportion of support (XY) and support(X), known as $\text{support}(XY)/\text{support}(X)$. In the example, confidence ($T \rightarrow C$) is 0.66, calculated from $\text{support}(TC)/\text{support}(T) = (2/5)/(3/5)$, which means if there is item T, the possibility that you can find item C is 0.66. Confidence can be understood as the estimation of conditional probability $P(C|T)$, which means the probability of the C's occurrence in those transaction under the condition that the transaction contain T's occurrence at same time (Jochen, Ulrich, Ntzer, & Gholamreza, 2000).

Lift is defined as the support (X|Y) divided by support(X) and indicates the quality of the rule. The quality indicates how reliable the mined rule is. In the example, if we want to know the lift of rule $T \rightarrow C$, the calculation

is as follows: the $\text{support}(T|C)$ refers to the support of T based on the existence of C , and there are 4 transactions containing T only two of them containing both T and C , so the support $(T|C)$ is $2/4$ in Figure 2.2, and $\text{support}(X)$ is $3/5$, hence $\text{lift}(T \rightarrow C)$ is $5/6$ which means there is 83.3% possibility that T occurs as a result of C 's occurrence.

Many algorithms have been used to discover association rules from spatial datasets. As Han & Fu (1995) proposed, meta-rules were used to find multiple level association rules. But the application of this algorithm makes meta rules re-designed, it is hard for users to control the rules. FP-growth algorithm was proposed by Jiawei Han based on FP Tree structure. The times of candidates generation and scanning database decrease by that algorithm, but it is more suitable for long-pattern extraction generation of candidates (Kondaveeti, Liu, Runger, & Rowe, 2011). DLT-Miner algorithm developed by Lee, Hong (2007) was applied to extract spatial association rules in image. Moreover, Agrawal and Srikant (1994) found an downward closure property named as Apriori, k -itemset is frequent only if all of its sub-itemsets are frequent, among frequent k -itemsets. It is one of algorithm used widely to find the association rules nowadays (Kondaveeti et al., 2011). It indicated that the frequent patterns will be mined by finding the frequent 1-itemsets via scanning the whole database, then using the itemset found previously to generate candidate frequent 2-itemsets, until no more frequent k -itemsets can be found (Koperski & Han, 1995), which is the original ideal of Apriori algorithm. In many cases, the Apriori algorithm efficient reduces the size of candidate sets using the Apriori principle, pruning techniques (Han, Cheng, Xin, & Yan, 2007) while guaranteeing completeness.

The Apriori algorithm find the frequent itemsets by generating candidate itemsets to compute support and prune the candidates each time by confidence from transaction database, and the lift was used to select the significant association rules (Ye & Chiang, 2006). The input of the Apriori algorithm is the minimum support and minimum confidence, the transaction dataset as well. The selection of interesting rules is based on measure of interest and significance (Hahsler, Grün, & Hornik, 2005). The minimum support and minimum confidence can be used to filter the interesting rules and lift gives indication for selecting the significant rules. Besides, to avoid the situation that too many rules are qualified with the minimum support and confidence, another interest measure lift can be used to further rank the rules.

Typically, when the rule's lift is high, its support is quite low (Hahsler & Chelluboina, 2011). So if the minimum support is too high, the rules with high lift tend to be filtered, otherwise, the rule with low support will be not useful for the application. Bayardo & Agrawal (1999) argued that the optimal rules often occur on the border of support and confidence. So the best minimum support can be chosen based on the scatter plot of confidence and support.

The Apriori algorithm consists of two steps: first, find the frequent itemsets, which are those that meet the minimum support condition, and then use the frequent itemsets to generate association rules. Then, the confidence of each association rule can be calculated and the rules that meet the minimum confidence condition will be selected as the results.

Consider Figure 2.2(a) as a transaction database, consisting of 9 transactions and in each transaction, it stores the corresponding items. In this example, the supposed minimum support is $2/9$ and the minimum confidence is defined as 70%.

Apriori scans the transaction database and generate a 1-itemset candidate. Figure 2.2(b) shows the 1-itemset and its support which are all more than $2/9$, so all the itemsets satisfy the minimum support value. Then, based on the itemsets that meet the minimum support in Figure 2.2(b), generate the 2-itemset frequent patterns by

join the itemsets in candidate table itself, referring to the combination of the items in 1-itemset table. 2-itemset and its support are shown in Figure 3.2(c), which indicates that the support of itemsets $\{A D\}$, $\{C D\}$, $\{D E\}$, $\{C E\}$ do not reach the minimum support. Therefore, these are eliminated from the 2-itemsets candidates. To generate the 3-itemsets frequent, combine the items that meet the support requirement in 2-itemset candidate table. For instance, in Figure 2.2, the 3-itemsets are $\{A B C\}$, $\{A B E\}$, $\{A C E\}$, $\{B C D\}$, $\{B C E\}$, $\{B D E\}$ generated from 2-itemset candidate table. However, according to the property of Apriori that all subsets of a frequent itemset have to be frequent, the 3-itemset $\{A C E\}$, $\{B C D\}$, $\{B C E\}$, $\{B D E\}$ will be removed, because 2-itemset $\{A D\}$, $\{C D\}$, $\{D E\}$, $\{C E\}$ are removed due to the not meeting the minimum support $2/9$. As a result, the 3-itemset $\{A B C\}$, $\{A B E\}$ are kept in the candidate table as shown in Figure 2.2(d). this process, removing the sub itemsets that do not meet minium support is called prune operation to get rid of heavy computation because of large candidate itemsets. Similarly, the 4-itemsets candidate $\{A B C E\}$ can be generated but its support is $1/9$ which is less than $2/9$. The maximum itemset depends on the number of genus type.

After the generation of frequent itemsets, the next step is to find the association rules. We regard all the candidate itemsets as a set including all the itemsets in 1-itemset ,2-itemset, 3-itemset, 4-itemset table and generate all nonempty subsets of from each frequent itemset. For each nonempty subset, the output rule will be derived from the subset of the frequent itemset from the set , then, the rule will be retained if its confidence meets the minimum confidence threshold. For example, from the itemsets table in Figure 2.2, a set refers to $\{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{A B\}, \{A C\}, \{A E\}, \{B C\}, \{B D\}, \{B E\}, \{A B C\}, \{A B E\}, \{A B C E\}\}$. One of the frequent itemset is $\{A B E\}$ from the set, so its subsets are $\{\{A\}, \{B\}, \{E\}, \{A B\}, \{A E\}, \{B E\}\}$. The possible rules can be derived based on the frequent itemset are as follow, $\{AB \rightarrow E\}$, $\{AE \rightarrow B\}$, $\{BE \rightarrow A\}$, $\{A \rightarrow BE\}$, $\{A \rightarrow BE\}$, $\{B \rightarrow AE\}$, $\{E \rightarrow AB\}$ and rules that meet the threshold will be the final mined rules.

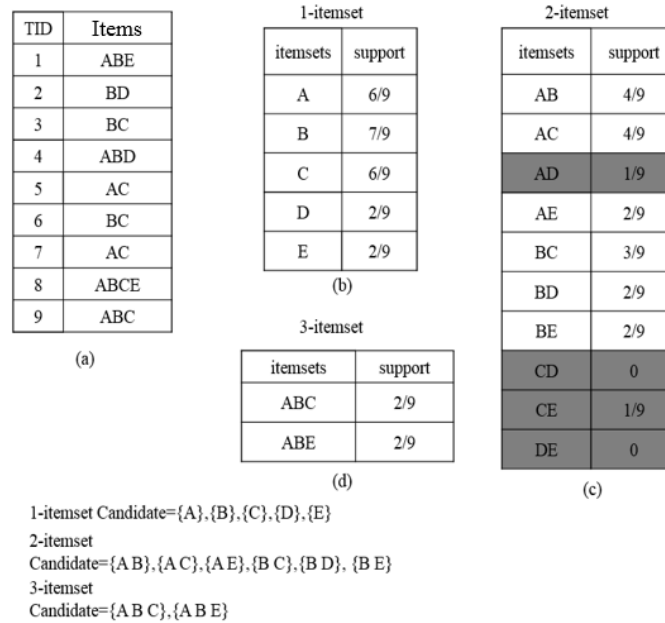


Figure 2.2 Apriori algorithm

2.2. Volunteered Phenological Observations

From the last decades, there is a notable increase in the interest in phenology, the study of the timing of life-cycle events of organisms (Van Vliet et al., 2003). The analysis of changes in the timing of phenological events contributes a lot to the study on potential influence of climate change on ecosystems (Inouye, Barr, Armitage, & Inouye, 2000). Besides, the research on phenological events spatially at the continental scale will be able to reflect the information and relationships among status of species, communities and ecosystem (New, 2005). And reliable phenological observations can be a useful resource to reveal and quantify large-scale patterns of species and its spatial synchrony/asynchrony in species response (Inouye et al., 2000).

Unfortunately, the incomplete nature of existing reported phenological observations and the lack of adequate constant observations in temporal dimension tend to impede the reveal on phenological variation characteristic in most situations (New, 2005). While the spatio-temporal database is incomplete, or inaccurate to some degree, it provides limited opportunities for making analysis spatially and temporally with some level of uniformity and completeness (Dobson, n.d.).

To obtain adequate phenological data, many phenological networks involving the volunteers' participation have been established worldwide. The common interaction between citizens and scientific research nowadays is data collection, considered to be the first step in wider scientific engagement. The phenological network allows volunteers to add a new record every time they see an animal or plant in phenological event, and the observations will store its geographic coordinates and time of sighting (Stodden, 2010). Even if phenological monitoring activities in the US are quite successful because of the integration with other ecological monitoring networks, remote-sensing products and data management capabilities, as well as the active participation by citizens and scientists, the sustainable phenological data in US is sparse (Schwartz, Betancourt, & Weltzin, 2012). To improve the use of phenological information in policy making as well as in research, there is an urgent need to expand research on improving the data collection protocol to improve the completeness and accuracy of the data and the lack of access to, and integration of data is partly caused by lack of information on limited datasets are available (Van Vliet et al., 2003).

2.3. Spatio-temporal Analysis in Volunteered Geographic Information

Many researchers have worked with volunteered observations in spatio-temporal analysis. In natural hazards domain, the disaster event once was extracted based on VGI Sensing which was used to define and monitor a geographic floods zone (Muelliganni, Janowicz, & Lee, 2011). And De Longueville (2010) has applied volunteered geographic information from social media Flickr to study how the human activity affects the urban growth in space and time. Besides, the extracation of intersting spatio-temporal pattern from VGI data also helped to explore and study the urban population characteristic (Aubrecht, Ungar, & Freire, 2011) . In the domain of phenology, the volunteered phenological observational data have been used to monitor complex ecological relationships, thereby tracking patterns such as phenology, geographic distributions and abundance of organisms (Crain, Cooper, & Dickinson, 2014). The National Phenological Network is a main data source of phenological observations collected by volunteers organized in regional or national phenological networks. One particular issue with the use of VGI is the question of data quality (Flanagin & Metzger, 2008). VGI research identified data quality as a major problem of community-based data acquisition and it is true in the scenario studied in the integration of geo-tagged reports about events or incidents (Yanenko & Schlieder, 2012). There are some previous studies justifying the use of NPN dataset. According to Mayer (2010), NPN has generated a standard protocol for phenological data collection to ensure the observations made by volunteers

reach the requirements of researchers' standard. The back-end strategies are used as well to ensure the store of the data that will be up to the quality standards (Mayer, 2010). Mehdi Poor et al. (2013) developed a geo - computational workflow to check the consistency of volunteered geographic information. Using the dataset from NPN to test the workflow, the result showed that the inconsistent VGI data are not wrong. Therefore, the volunteered phenological observations from NPN assumed sufficient for the study.

3. MATERIALS AND METHOD

3.1. Materials

USA National Phenology Network offers observations, collected by volunteers, to support scientists to get more insights in the relationship between plant and animal phenological changes in order to support decision making (USA National Phenology Network, 2011). Phenological observations used in my research were downloaded from the this website for the period 2008 to 2014, containing 3,447,015 records about 600 species. The elevation range for all the observations is from 0 to 3880 meters. Although, the website offers the data from 2008 to present, the data from 2015 is not complete for the whole year and was ignored. The downloaded phenological observations only concerns species of plants, because animals tend to move over time, the locations change for each creature.

USA-NPN offers two types of data: raw status data and summarized data. Raw status data offers information about plant or animal species's observed location and the date for its status of phenophase, for example, latitude, longitude, elevation, phenophase description, species name, genus name, DOY and so on. "DOY" refers to the day of year when one phenological event is observed. Besides, an attribute of phenophase status is offered. When the value is "yes" it indicates the phenophase is occurring, otherwise, the value is "no". Observations are also offered as summarized data (Table 3.1). Like raw, these data also contain geographic information and status data for each observation. Besides, each row contains detailed information about the phenophase duration by summarizing the records based on the phenophase status in raw status data. This data type gives support to phenophase onset, duration and end estimation (USA National Phenology Network, 2011). Each record in summarized data, the date of first "yes" and the date of last "yes" was summarized from the raw status data. Both the duration of the phenophase and the geographic information are needed in the research, the observations are used from summarized data.

Table 3.1 Records showing instances in volunteer phenological observation data from USA-NPN

Latitude	Longitude	Elevation in meters	Species name	Genus name	Phenophase description	The DOY of first "yes"	The DOY of last "yes"
35.214489	-97.472557	350	common lilac	Syringa	Open flowers	100	100
43.08535	-70.69133	12	red maple	Acer	Breaking leaf buds	125	132
43.08535	-70.69133	12	sugar maple	Acer	Falling leaves	276	305

3.2. Method

This section presents the methods applied to realize the mining of co-location patterns in space and time to complete the research. An overview of the steps is provided in Figure 3.1.

First of all, we removed incomplete observations in the phenological database and selected the species or genus based on the defined criterias. Then, the data were extracted respectively for spatial co-location mining based on distance, distance and elevation and for spatio-temporal co-location mining based on distance and relative time interval. Secondly, two kinds of transaction data was prepared for spatial and spatio-temporal co-location mining spatial. Spatial transaction data was generated based on distance only, and on distance and elevation. For the spatio-temporal transaction data, we created the transaction data by distance and relative time. Thirdly, both spatial co-location rules and spatio-temporal co-location rules were mined by applying the Apriori algorithm to the transaction data. Valuable co-location rules were identified by using measures, confidence, support and lift for different size of neighborhood. As a result, both spatial and spatio-temporal co-location rules were presented in graphs and maps.

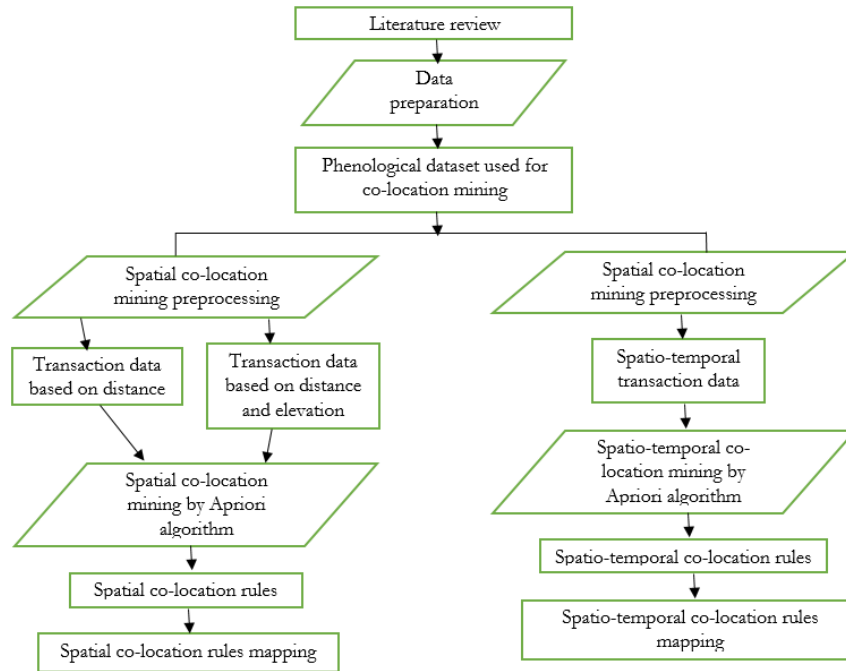


Figure 3.1 Method workflow to answer the research questions

3.2.1. Data Preparation

We used the dataset from the summerized data, because the information of phenophase duration was needed. In the spatial co-location mining process, only coordinates information (latitude, longitude, elevation) and the species name were needed. But in the spatio-temporal co-location mining process, the time range for each observations' phenophase occurrence was needed to obtain the species pairs.

The first step of data preparation was to clean the incomplete dataset, which means removing the phenological records with "0" value in geographic coordinates and with "-9999" in elevation. The raw data obtained from the dataset is in the format of CSV for different years. To manipulate the data in an efficient way, the phenological data from 2008 to 2014 were organized in PostGIS. The data per year was stored as one table in the database.

The summarized data offers latitude and longitude only for genus, but in the following step, the distance calculation among different locations will be needed to define the region of the co-located neighborhood for multiple genus. Therefore it is convenient for distance calculation to convert geographic coordinates to geocentric. The main concern about the projection selection is to represent the distance reasonably well and the study area is the continental United States, and U.S. National Atlas Equal Area projection presents the characteristics, wider east to west than north to south, in an optimized way (Corey, 2013). Therefore, the EPSG 2163 was chosen to convert the latitude and longitude into meters.

The second step was to select the species used for co-location mining. The general criteria for species selection are as follows: each species should have a minimum number of observations; the chosen species must have enough unique observational locations; the selected species should be those that have records along all years. However, the data resource is from volunteers, the data is not perfect and does not meet all the requirements, so it is hard to define an explicit minimum value for the number of observations or unique locations. Moreover, both research objectives are concerned with the co-location patterns in space. Thus, the distribution of the observations for different species. The attribute genus indicates a general grouping of the species. For example, Quercus genus includes both deciduous plants, like black oak, northern oak, white oak, and evergreen plants, like laurel oak. The distribution of deciduous species in Quercus were concentrated on the northern part of the US, and the evergreen species in Quercus were distributed in the southern US. So when using the original genus to mine the spatial co-location pattern, some important pattern may be missed because of the way species are grouped. To solve this problem, we made a sub grouping under the genus type. The Quercus was divided into two categories. The species laurel oak and live oak in the dataset, the genus name were changed as Quercus1, referring to the deciduous plants in Quercus. And the rest of the species in Quercus were stored as Quercus2 referring to the evergreen plant.

To make full use of the available data and select the appropriate species, exploratory data analysis were done. To select the species that occur each year, a new attribute, number of years was added to each table. This attribute indicates the number of years in which the species occurred. So the species could be selected when the value of number of year was 7, indicating it occurred from 2008 to 2014. Due to the fact that the number of unique locations and the observations for each species affect the number of transactions, the number of unique locations for each species were calculated to decide which species to choose.

In the third step, two phenological tables were prepared for analysis of co-locations. For the spatial co-location mining, the spatial co-location mining table was formed by the attributes: the unique locations' coordinates, elevation and the selected species name. For the spatio-temporal co-location mining, the table was composed

of the attributes: unique locations' coordinates, phenophase description, species name and the DOY of first "yes", the DOY of the last "yes". The time range for specific phenophase for the species under the same genus was aggregated because species under one genus have the similar phenophase. Each observation has the data of the DOYs of the first and last yeses. Genus was used as the unit for co-location mining process. The DOYs of first and last yeses for multiple species under one genus was aggregated. For example, one record of the time range for DOYs of the first and last yeses for sugar maple's falling leaves is from 212 to 309, and the time range for red maple's falling leaves is from 239 to 323. Therefore, these two kinds of species have overlapping time range for falling leaves, and the time range for *Acer*'s falling leaves is stored as from 212 to 309. The aggregation for the time range was that selecting the minimum DOY of the first yes, and the maximum DOY of the last yes from the species records.

3.2.2. Spatial Co-location Mining

3.2.2.1. Spatial Co-location Mining Data Preprocessing

The Apriori method was applied to mine the spatial co-locations. This method identifies species that appear nearby by mining frequent patterns from a set of transactions (i.e., {TID: itemset}), where TID is a transaction id and itemset is the set of items included in transaction TID (Han et al., 2007). Generally, in association mining problems, a transaction in a database contains a transaction ID and many distinct items. The number of distinct items will be regarded as an item set (Hahsler et al., 2005). In my research, each unique location is considered to be a transaction and the corresponding items in each transaction are the different species occurring in that location. So the transaction data will show the unique locations and the species that occur in that location.

Spatial patterns refers not only to the relationship among the genus that occurred at a common location, but also the relationship among the species happening in different but nearby locations (Shekhar & Huang, 2001). To quantify the concept of "nearby", a way to define the neighborhood should be introduced for a given neighborhood relationship. The neighborhood was defined in two ways: 1/ based on Euclidean distance and 2/ based on both distance and elevation. By Euclidean distance.. regard the locations within a given value of distance as located in the same neighborhood. The distance between one location and the rest of the locations will be calculated. We grouped the location and other locations of which distances are less than a user-defined value as one transaction. The related item set to the transaction contains all the species that occur at the grouped locations. To group the transaction data into neighborhood transaction data, we used the coordinate value of each transaction and calculate the distance among one location and the rest of locations. Based on the user-specified threshold for the distance, the spatial transaction data based on distance will be generated, which means all the locations that are within the minimum distance will be grouped as one neighborhood. When grouping the transactions based on distance, one pattern involving the same location points may exist in multiple neighborhood transaction, so the pattern formed at the same locations would only be grouped in one transaction to avoid the enhancement for the pattern. For neighborhoods defined both on distance and the elevation, elevation tends to influence the species meso-scales (Phillips, Anderson, 2005). Elevation is also a factor that can exert the diversity of species and affect the species on phenological stages (Premoli & Raffaele, 2007). Thus, understanding the effects of elevation to the spatial co-location pattern is necessary. As mentioned before, the distance was the only parameter to define the size of neighborhood. However, because of the elevation range, the species can distribute in the same location but with different elevation, which can also be regarded as spatial co-location. Similarly, the creation of transaction data was the same as the creation based on distance only. After grouping the transaction based on distance, each group would contain multiple species that occur in the neighborhood. And the species in each group had different value of elevation. In each group, the elevation difference would be calculated among different species. So the species pairs would be generated

if two species' elevation difference is less than a user-defined value. The value depends on the distribution of species, which would be experimented with depending on the dataset. The spatial transaction data based on distance and elevation would be in the format that each row represents a neighborhood which was developed based on distance and the species pairs falling in each row indicate the species that have co-location in elevation dimension.

3.2.2.2. Spatial Co-location Rules

After preparation of the spatial transaction data, spatial co-location were mined using Apriori. The rules depend on the following parameters defined by the user: size of the neighborhood (distance), minimum support and minimum confidence. Several distances were tested by changing the radius of the neighborhood, minimum support and minimum confidence.

A large set of rules can be obtained as the output of the algorithm, but only few rules with high lift and high confidence are valuable ones. So the parameters, minimum support and minimum confidence help to filter out useless rules. The rules with high lift tend to have a relatively lower support. According to Bayardo, Jr. and Agrawal (1999)'s research, the valuable rules are likely to exist on the support and confidence border. If the minimum support is too small, interesting patterns cannot be found easily; If the minimum support is too high, the rules with high lift will be filtered. So the definition of an appropriate support and confidence is important for obtaining useful rules. To find the appropriate minimum pre-defined threshold, it started from the overview of the support for each species existing in the spatial transaction data. The histogram of support for species that were ordered by the support gave an indication of the minimum support value. The median value of the support always tends to be a good start of the experiment.

It is difficult to select the minimum threshold only by looking at the series of number. So a scatter plot with support and confidence on the axes, the shading of the dot representing the lift can make all the rules and its measures visualized. From the scatter plot, the border can be seen clearly. Thus, the appropriate minimum support and confidence can be tested based on the confidence and support border.

The mined rules are provided in a text format that lists the involved species and the lift and support of each rule. Hahsler (2011) proposed a new visualization technique to show association rules and implemented them in the R package called ArulesViz. Using graphics for rules presentation makes it easy to find and explore the importance of the rules. Graph-based visualization uses vertices and arrows to show the association rules, where the vertices represent the involved items and the arrows represent the relationship among items (Hahsler,2011). The interest measures of each rule are presented via the color and width of the arrows. When using graph-based techniques to show a large set of association rules, it is easily to get cluttered (Hahsler & Chelluboina, 2011). After the filtering by the minimum threshold, the final result for spatial co-location rules with high lift form a very small set which can be visualized by graphs.

There are two kinds of graph-based visualizations based on vertices. One of it uses items as vertices. This visualization mainly concentrates on the composition (individual items) involved in each rule and it can make the rules that share items very obvious. Arrows are used to show the relationship between items and the value of lift, and the size of circles in the middle of arrows indicates the support value. In spatial co-location rules, the items refer to the species. The arrow indicates the relationship of which species tend to co-locate together and the lift for that rule. Another method is that the item sets were regarded as vertices. Each rule was formed by item set in the bracket and the arrows indicating the relationship. The width of the arrow presents the value

of lift and the shading of arrows indicates the value of support. All the rules can be seen clearly in this way and the item sets sharing different rules can be presented at the same time.

All the spatial co-location rules were generated based on the locations so it is necessary to visualize the rules in the map to get a better understanding of the species distribution. Each spatial co-location rule was presented in a map. Because the rules were mined based on grouping geographic coordinated within a neighbourhood, for all the phenological observation sites, a circle buffer was generated to present the size of the neighbourhood based on the distance value. The area within the circle indicates the district where the spatial co-location rule is likely to happen.

3.2.3. Spatio-temporal Co-location Mining

The spatio-temporal co-location mining was to find the species that locate nearby and have the same timing of the phenophase. To apply the Apriori algorithm to obtain the spatio-temporal co-location rules, the first step was to create the spatio-temporal transaction data.

Spatio-temporal transaction data was generated from the spatio-temporal co-location mining table. The process from transaction data generation to grouping the transaction based on distance was the same as the spatial co-location mining process. Each observation had the information of coordinates, phenophase description and the time range for the phenophase. After the spatial grouping, we combined the species for which the range of observed timings of a given phenophase overlaps more than 2 days. This combined item was stored as a grouped transaction. In this process, we used the relative time between two species instead of grouping the time range into different time interval. If we would have generated the transaction based on the time interval, species that have a similar timing of phenophase but the timing are closed to the end point of the time range would be separated into two different time interval. In that way, species that tend to occur in the similar time would be ignored. Thus, the better way to avoid this situation was to create the species pairs based on the relative time interval overlapping. The spatio-temporal transaction data would be in the format that each transaction contains multiple species pairs. The transactions refer to different neighborhoods generated by distance and species pairs refer to the species that tend to occur together within 2 days. In spatio-temporal transaction data, the genus pairs, two genus that have the overlapping time range for specific phenophase were stored in each neighborhood. When combine the genus into pairs, the time range for each genus has a corresponding phenophase for the DOYs of the first and last yeses. After the acquisition of the co-location rules, it can reflect the most frequent occurring genus and its phenophase in transaction data.

After the preparation of spatio-temporal transaction data, the process of mining spatio-temporal co-location rules was the same as describe in section 3.2.2.2. The configuration of the parameters was crucial to obtain the valuable co-location rules. Output of the algorithm are the spatio-temporal co-location rules with a high lift. The measures, confidence, support and lift for each rule were obtained as well. Each rule was composed of several species pairs which means they tend to occur at the same time and located near each other. Therefore, the species in the species pair are the species that have similar timing of phenophase. In addition, the time range for the species that have the similar timing of phenophase and the phenophase can be extracted.

The spatio-temporal co-location rules were presented using a graph-based technique. Apart from mapping each rule in a map, a graph would be created, to show the temporal element of the each rule. Related to each map, the graph includes a timeline showing the day of the year for the phenophase on the x-axis and using the value “1” and “0” as y-axis to show the species’ phenological event existence situation at the specific day. The time

range for the similar timing of phenophase and the phenophase can be seen clearly in the graphs. Hence, the spatial and temporal characteristic from the co-location rules were expressed.

4. RESULT AND DISCUSSION

4.1. Data Preparation

The USA-NPN data records were checked for completeness. This analysis identified and removed 233 records that had incomplete coordinates (i.e. had “0” or “9999” values). Exploratory data analysis was used to select the appropriate species for mining co-locations. Figure 4.1 presents the total number of observations per year. The total number of observations in 2008 is significantly smaller than for the rest of the years. Figure 4.2 illustrates the total number of unique locations per year and, again, 2008 has a very low number (165 compared with more than 500 for other years). The unique number of locations should meet the criteria mentioned in chapter 3 for mining co-location patterns because the preparation for transactional database was created by the unique locations. The amount of locations affects the amount of the transactions. Therefore, the data of 2008 was removed from the phenological database.

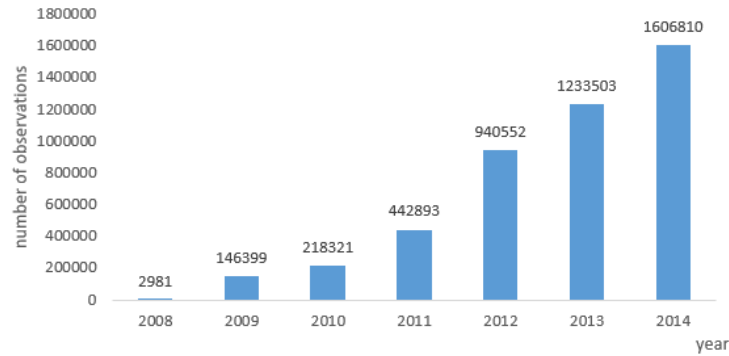


Figure 4.1 Histogram of the observations available for each year

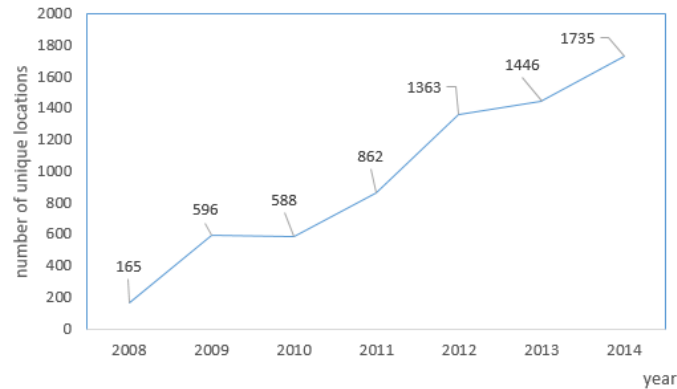


Figure 4.2 Total number of unique locations for each year

Figure 4.3 presents the total number of species and genus types observed each year. We found that more than 500 kinds of species were monitored between 2012 and 2014, even the year 2009 has more than 150 kinds of species. Too many types of species leads to a sparse distribution of observations and this negatively affects the result of co-location pattern mining. In order to get more unique locations for each items, considering multiple

species may have similar phenophase and characteristics, we decided to use the genus type instead of species to mine the spatial and spatio-temporal rules. The average number of genus type for each observation site is 4, which shows that at the same location, around 4 kinds of species were observed for its phenological event. Only single species' information was recorded and observed at 39.48% locations of all. 14.72% locations have the observations concerning 2 species type. It reflects that more than 60 percent of the locations have multiple species' observations. And each observation includes the person ID which indicates who, the volunteer, collect the observation. Multiple volunteers report phenological observations at the same site.

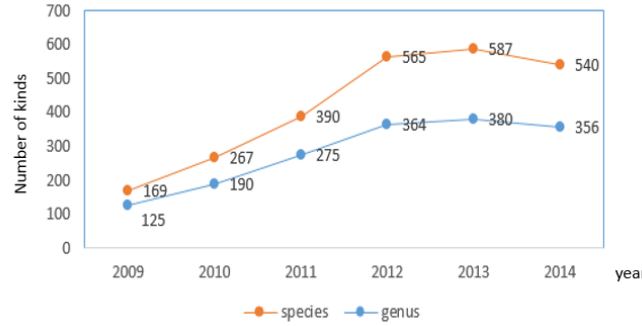


Figure 4.3 Total number of genus types and species types for each year

There were 54 genus that have records all the year and that all of them were used for the analysis because filtering them would result in a dataset with very few locations. Few locations means that few co-location patterns can be found so, the rules will not be interesting. The spatial co-location mining table and spatio-temporal mining table were created for the transaction data generation. The observations of 54 genus from 2009 to 2014 were extracted. And the extraction of the attributes latitude, longitude, elevation and genus name from the database formed the spatial co-location mining table. And spatio-temporal co-location mining table are the observations with attributes: latitude, longitude, phenophase description, species name and the DOY of first "yes", the DOY of the last "yes".

4.2. Spatial Co-location Mining

4.2.1. Spatial Co-location Mining Preprocessing

After the data preparation, we created the transactional data based on the phenological dataset so as to apply the Apriori algorithm to mine the spatial co-location rules in two ways.

Firstly, we prepared the spatial transaction data based on distance. The spatial co-location mining table had 5660 records because the same coordinates may occur many times in transaction data as different genus may occur in the same geographic location. After coordinates conversion, each unique location was considered to be a transaction and the corresponding items in each transaction are the different genus occurred in that location. The transaction data showed the coordinates and its occurring genus. Figure 4.4(a) presents a sample of the transaction data. This data consists of 2351 rows generated from the 5660 observations described above. The reduction in the number of transactions is caused by the fact that several genus can occur in one geographic location. For example, there are 3 different type of genus in the 4th transaction, which means that 3 types of genus, *Bombus*, *Cornus*, *Quercus* occurred in that location. Spatial transaction data were generated based on distance, the radius value from 0km to 40km, the interval of that is 5km. Each transaction was grouped based on radius value and the genus were combined as well to form the neighbourhood transaction data. In Figure

4.4(b) we can see that the grouping of locations may cause duplicated groups. In this example, the two points in the third and the fourth groups are actually the same. Therefore, duplicated groups were removed. This grouping process further reduced the number of transactions. For example, for a neighbourhood radius of 5km, the spatial transaction data contains 1531 rows, which means that around 800 locations were grouped into other transactions. The spatial transaction data was created by extracting the genus name from each transaction to the grouped transaction. As shown in Figure 4.4, each group contained more than one coordinate, and search the corresponding coordinates in transaction data to get the value of genus and store them into the neighbourhood transaction data. Different neighbourhoods may occur the same type of genus that will make the items duplicated. To prevent this only the unique genus are added to the final transaction database, as shown in Figure 4.4 (c).

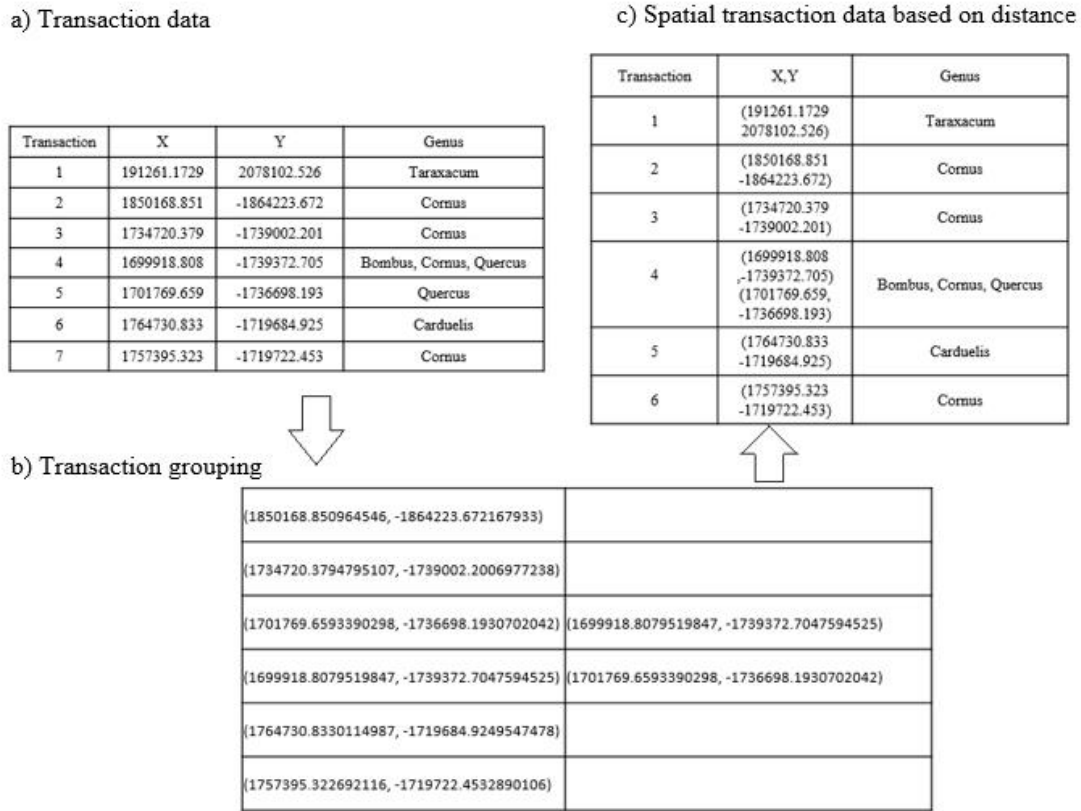


Figure 4.4 Process from transaction data to spatial transaction data (Sample dataset)

Secondly, we generated the spatial co-location transaction data based on both distance and elevation. Therefore, the elevation attributes from the spatial co-location mining table were also used to create the transaction data. The radius for the neighbourhood generation were tested from 0km, 5km, 10km until 40km. The distance interval was set as 5km to create multiple spatial transaction data. When the distance interval was very small, like 2 km, the genus within a neighborhood were quite the same and the support for the rules mined were nearly the same. So using 5 km was better to see the different rules from different size of the neighborhood.

After the grouping process based on distance. Different elevation difference were tested to generate the transaction data. The elevation differences from 30 meters to 200 meters were tested. The 100 meters was the best one because the genus pairs start to change notably from the distance-based transaction data. When the

elevation difference was set to 100 meters, the genus pairs generated within in the same elevation difference. Figure 4.5 presents the sample result of spatial neighborhood transaction data based on both distance and elevation, two coordinates were grouped as one transaction in the 4th transaction, and because the elevation difference among *Bombus*, *Cornus* and *Quercus* was less than 30m, these 3 types of genus were combined two by two as species pairs.

NTID	(X,Y)	Genus
1	(191261.1729,2078102.526)	<i>Taraxacum</i>
2	(1850168.851,-1864223.672)	<i>Cornus</i>
3	(1734720.379,-1739002.201)	<i>Cornus</i>
4	(1699918.808,-1739372.705) (1701769.659,-1736698.193)	<i>Bombus</i> + <i>Cornus</i> , <i>Cornus</i> + <i>Quercus</i> , <i>Quercus</i> + <i>Bombus</i> , <i>Carduelis</i>

Figure 4.5 Spatial neighbourhood transaction data based on distance and elevation

4.2.2. Spatial Co-location Rules

Spatial co-location rules were mined by the Apriori algorithm. The spatial transaction data was used as the input of this data mining algorithm.

Firstly, we mined the spatial co-location rules by the distance spatial transaction data. Minimum support and minimum confidence were tested to find valuable rules. Choosing the best minimum support and minimum confidence for each size of neighborhood directly affects rules we found. When the minimum support and minimum confidence was small, more than 10 thousand rules were generated. Figure 4.6 shows a scatter plots of spatial co-location rules and its support, confidence and lift. These rules were mined using the various transaction data generated by varying the radius of the neighbourhood. These scatter plots were generated using low minimum support and confidence values, therefore, many rules can be seen in the graph.

We found that increasing the radius of the neighbourhood leads to more rules. When the radius was 0km, only 356 rules were mined. When the radius reaches 35km, more than 440,000 rules were mined. The increase of the size of the neighbourhood makes more kinds of genus fallen into one neighbourhood so more frequent genus items can be found. The support and confidence are presented on the axes and the lift is shown with the shading of the dots in Figure 4.6. The darkest dots tend to appear close to y axis indicating that the rules with high lift were always with lower support but quite higher confidence. Besides, the dots pattern become larger and wider to the right from radius 0km to 40km. This suggests that the number of rules with high support increases. All the information offers indication for the selection of minimum threshold that the minimum support would increase as the size of neighbourhood increased. The rules with high lift which were the darker dots from the graph, always exist when the confidence is more than 0.6. So the minimum confidence was set as 0.6. The minimum support was fixed by analysing the support histogram for all genus. the median support obtained from the support histogram are shown in Table 4.1(a). The median support for all genus increased as the neighborhood increased. Thus, median support value were the starting value to filter the spatial co-location rules.

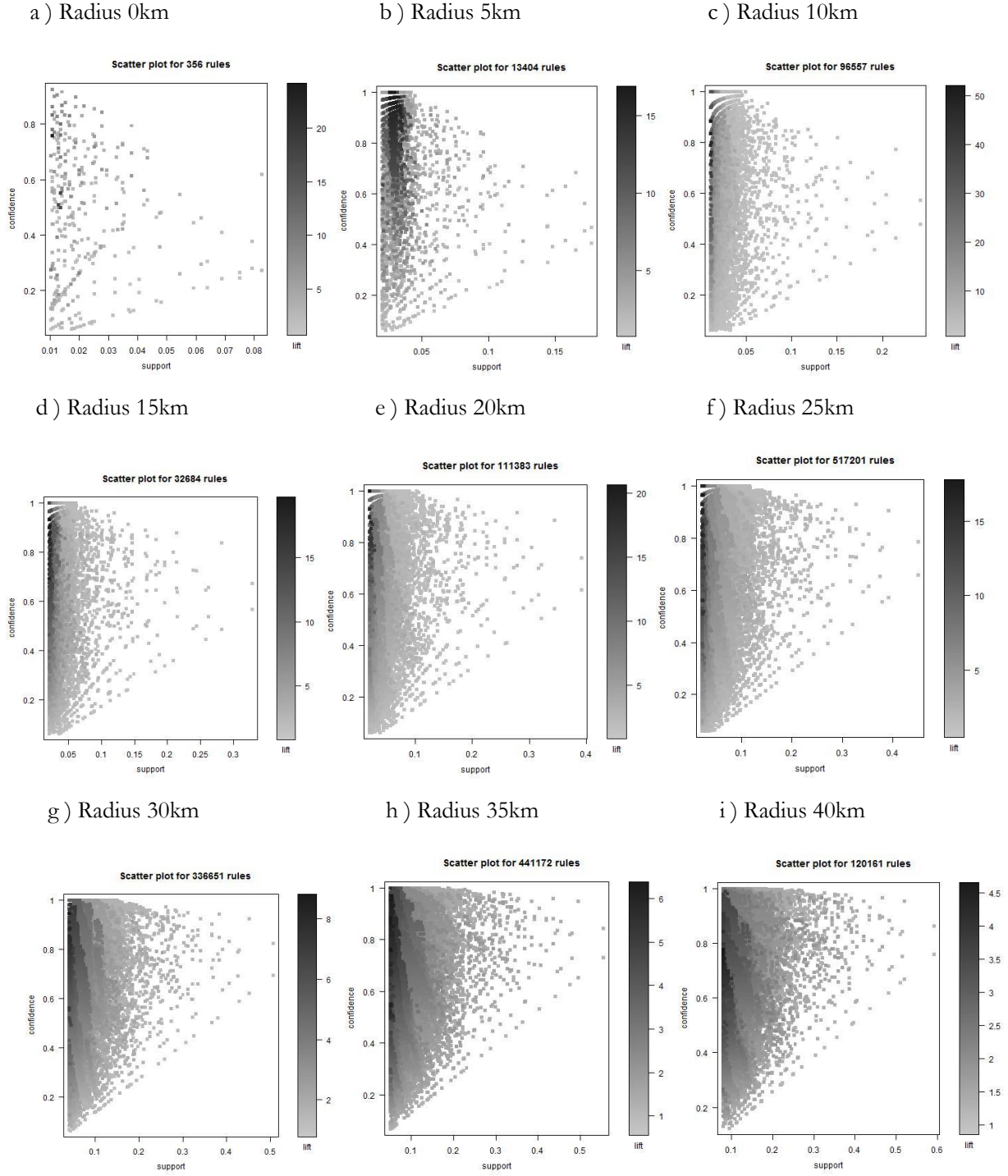


Figure 4.6 Scatter plot of spatial co-location rules for different size of neighbourhood (Distance)

Table 4.1 Median support derived from support histogram shown in Table (a) and the best minimum support for the spatial co-location rules based on distance shown in Table (b)

a)		b)	
Distance/km	Median support	Distance/km	Min_support
0	0.03	0	0.04
5	0.05	5	0.12
10	0.10	10	0.18
15	0.18	15	0.25
20	0.25	20	0.32
25	0.30	25	0.38
30	0.36	30	0.45
35	0.40	35	0.48
40	0.46	40	0.50

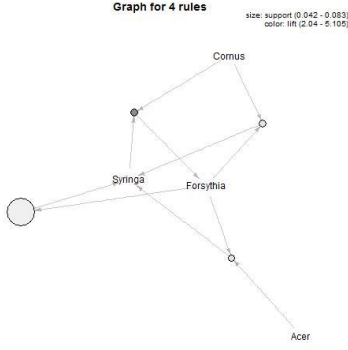
The valuable rules are those with high lift and with confidence and support values relatively high. The best minimum support are shown in Table 4.1(b). Taking the result generated from the neighborhood with 20km radius as an example, 3 spatial co-location rules were obtained when the minimum support was 0.32 and confidence was 0.6. These values mean that in 32% of the occasions these genus were together and that when one of the genus is observed in one location, the possibility to find another genus in its neighborhood is of 60%. The lift value shows the quality of the rules. Rule 1 was that Forsythias and Syringa tend to co-locate with each other with lift 1.392. Rule 2 refers to Acer and Syringa's co-location with lift 1.162 and Third rule with lift 1.12 are concerning Cornus and Syringa. From different size of neighbourhood, at least 3 or 4 rules were selected with high lift and high minimum support and confidence. The most frequent rule from different size of neighborhood, was that Acer and Syringa tend to occur nearby. We may doubt that the mined co-location rules were just presenting the most common genus in phenological dataset. Actually, Syringa, Forsythia are quite common in the dataset but Cornus was not with many observations in the dataset. And many other type of genus like Carya, Fagus which have more observations were not involved in the co-location rules. So the spatial co-location rules do show something different and interesting from the large spatial dataset.

As the result obtained from the Apriori algorithm were in the format text, it was trivial to read the rules one by one from the text only. So the spatial co-location rules were visualized to explore the information to grasp the mined rule much better than a text file.

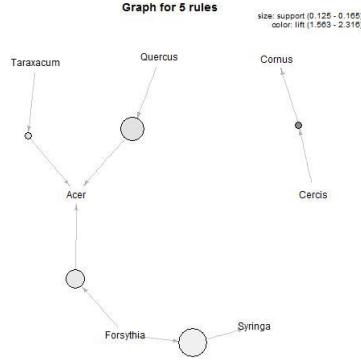
Spatial co-location rules generated from distance-based transaction data were presented by graph-based visualization technique using items (genus) as vertices. All the co-location rules in different size of neighborhood are visualized as Figure 4.7, which presents the co-location relationship among genus. Taking Figure 4.7 (e) as an example, when the radius is 20km, this graph clearly presents Acer, Forsythia, Syringa and Cornus were the involved genus type in the top 4 rules. And the rule that Forsythia and Syringa tend to locate nearby was the strongest rule because the circles' shading referring to lift was the darkest one, and the size of the circle which means the support were not too small. As for the co-location relationship between Acer and Syringa. The support for this rule was the highest one but lift was lower than the previous rule. In addition, Syringa was at the center of the graph which means that mined co-location rules were both related to this genus. Thus Syringa was the most possible genus type to co-locate with other genus. Forsythia, Cornus, Syringa are the most frequent genus type repeatedly occurring in all the size of neighborhood results. And Syringa and

Forsythia are the two genus of 54 genus that involved many rules in each size of neighborhood. It indicates that Forsythia and Syringa tend to co-locate in space.

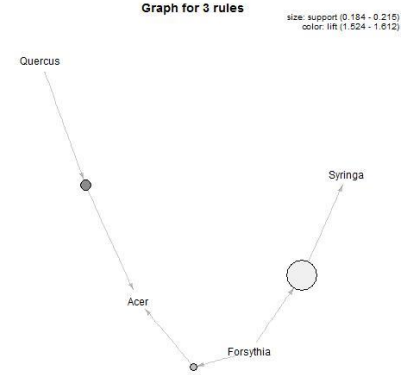
a) Radius 0km



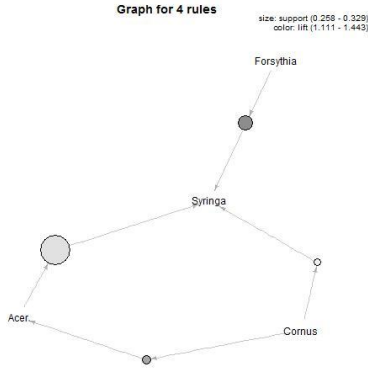
b) Radius 5km



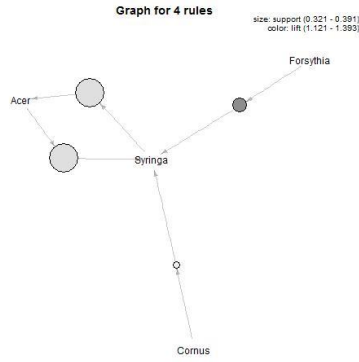
c) Radius 10km



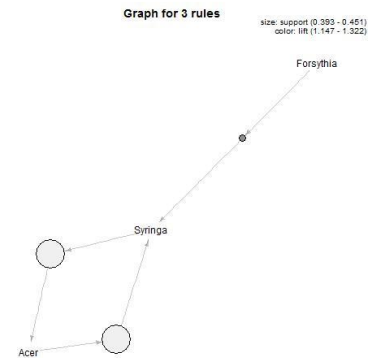
d) Radius 15km



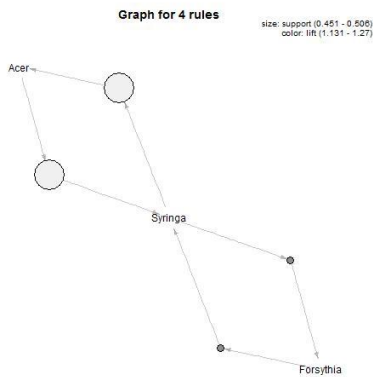
e) Radius 20km



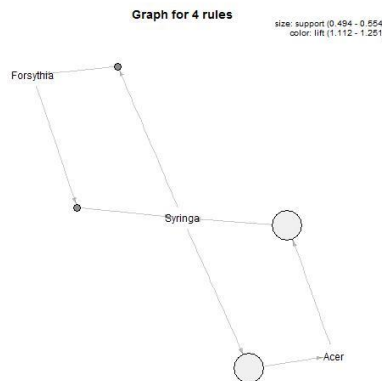
f) Radius 25km



g) Radius 30km



h) Radius 35km



i) Radius 40km

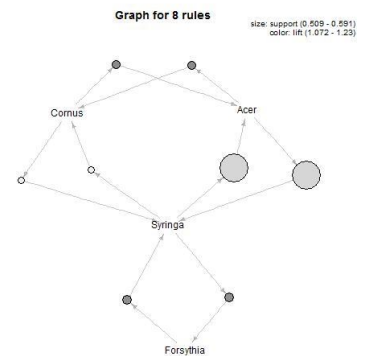


Figure 4.7 Graph-based spatial co-location based on distance rules visualization using items as vertices for different neighborhood size

Figure 4.8 shows 3 top rules with high interests measures in 3 maps when the neighborhood size was 20km. The size of the circles indicated the possible area where the co-location rules can happen. Syringa was the genus involved in all the maps. And Syringa, Cornus, Acer and Cornus distributed densely at the north-eastern part of the US, especially the eastern coast, so the states Maine, Connecticut, Massachusetts and New Jersey were covered by the co-location rules area. Florida were the state that Cornus and Syringa tend to co-locate with each other.

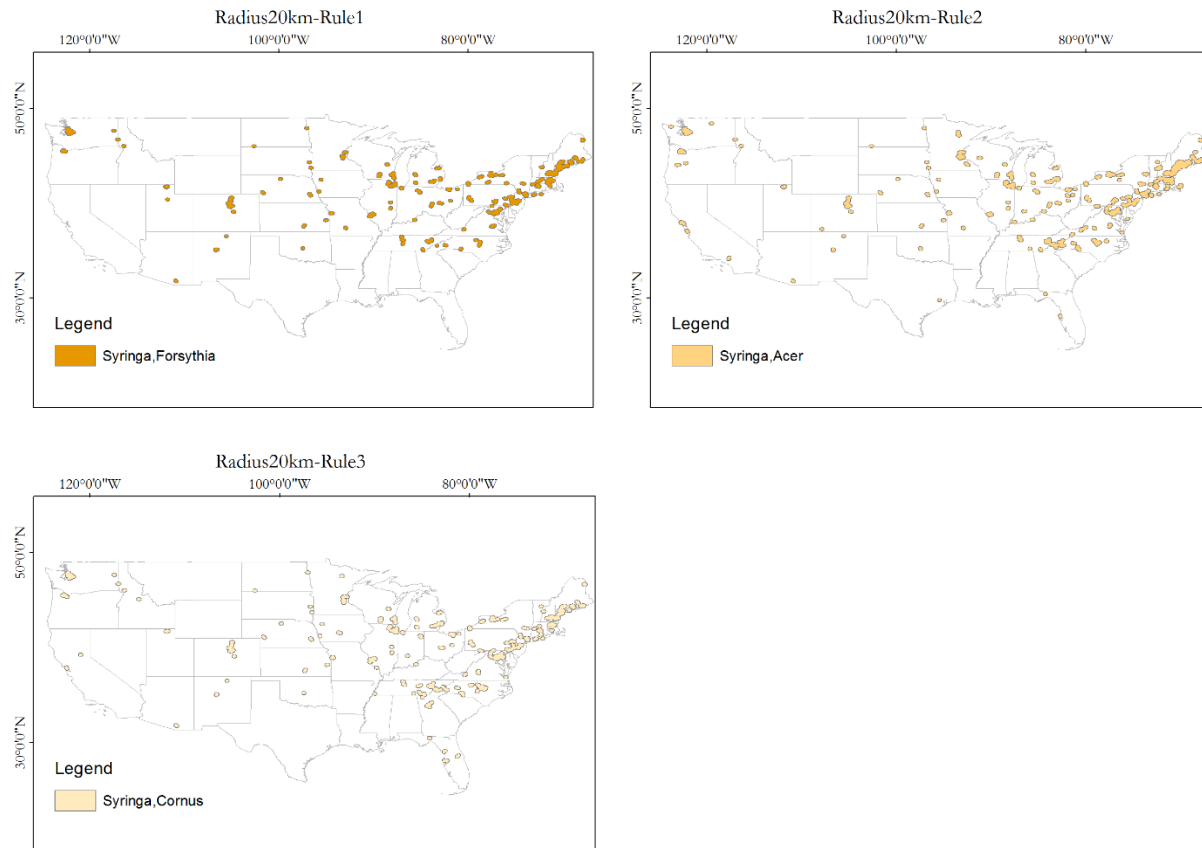


Figure 4.8 Spatial co-location rules map generated by distance when the distance is 20km

Secondly, due to the elevation can affect the genus distribution and the genus timing of phenophase, we applied Apriori algorithm into the spatial transaction data based on distance and elevation. We configured the radius value the same as the previous to generate spatial transaction data by distance and elevation. The genus were combined when the elevation difference was less than 100 meters. In general, considering elevation reduces the the minimum support with respect to the previous case. The combination of different species from elevation differences increases the number of items in spatial transaction data, leading to the decrease of support for the genus pairs co-occurrences.

Figure 4.9 shows the scatterplots of the rules for the case when distance and elevation are used to define the neighborhood. In this figure we see that the lift range for these rules is much larger than the ones found when only using the distance-based (Figure 4.9). But the support range for different size of neighborhood, generally it was quite low. Most values were smaller than 0.2, which shows that the distribution of the co-occurrence for multiple species pairs is really sparsely. Moreover, the darkest dot that indicate the high lift from the graphs

exists above 0.5. therefore, the minimum confidence were set as 0.5. Based on the support histogram, the median support (Table 4.2 (a)) for different size of neighborhood transaction were used as the initial threshold to find the rules. As the radius become larger, the number of transactions decreased and the number of transactions where the genus pairs increased. There was the slight increase for support .

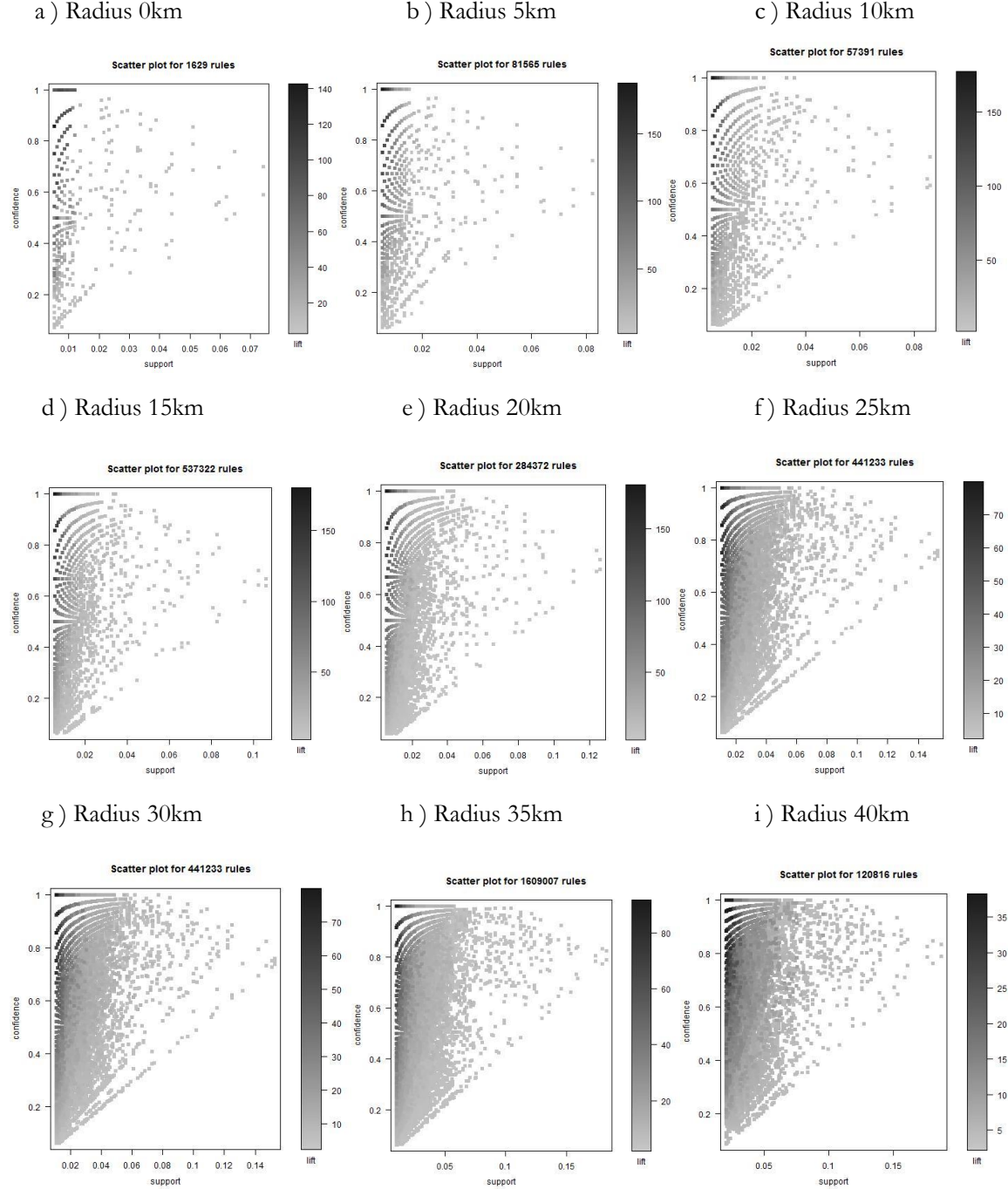


Figure 4.9 Scatter plot of spatial co-location rules for different size of neighborhood (Distance and elevation)

Table 4.2 Median support derived from support histogram shown in Table (a) and the best minimum support for the spatial co-location rules based on distance and elevation shown in Table (b)

a)

Distance/km	Median_support
0	0.03
5	0.04
10	0.06
15	0.08
20	0.10
25	0.12
30	0.13
35	0.16
40	0.16

b)

Distance/km	min_support
0	0.045
5	0.058
10	0.068
15	0.09
20	0.12
25	0.15
30	0.152
35	0.174
40	0.180

Based on many experiments, the final minimum support used to mine the elevation-based spatial co-location rules are shown in Table 4.2 (b). The spatial co-location results showed the genus that co-locate with each other in both distance and elevation. When the distance was 0km, the rule with highest lift was that Syringa, Cornus, Cercis and Forsythia tend to co-locate nearby. The support for this rule was 0.05, the confidence was 0.68 and the lift was 8.03. When the distance reached to 20km, the rule that Taraxacum, Syringa and Acer tend to occur nearby and within 100 meters in the height was the rule with highest lift 7.9. We found that the spatial co-location rule mined with high lift when the distance was 0, did not exist in the result when the size of the neighborhood was 30km. It indicates that the increase of the neighborhood size makes the minimum threshold increasing. And the rule mined in smaller size of neighborhood had lower support value than in a larger neighborhood. In addition, Syringa and Forsythia are the genus that frequently appear in the rules mined in each size of the neighborhood, showing that Syringa and Forsythia tend to co-locate in space at a large possibility no matter what the size of the neighborhood is. And within the same size of neighborhood, several rules were mined. Syringa, Forsythia, Acer and Cornus appear very often in all the rules.

In order to grasp the rules easily, the spatial co-location rules mined from distance and elevation based spatial transaction data are visualized in two ways in Figure 4.10 and Figure 4.11. Taking the graph when the neighbourhood is 20km as an example, in Figure 4.10 (e) and Figure 4.11(e), Forsythia and Syringa, Syringa and Cornus co-located in elevation dimension, and the Syringa, Forsythia and Acer tend to co-locate in space with each other. So was Cornus, Syringa and Forsythia. Each rule and its support and lift were presented in Figure 4.10. In Figure(e), two rules between two genus pair Acer and Syringa, Syringa and Forsythia are shown, but in Figure 4.11(e) there is only one arrow shows that two genus pairs, the rules without duplication can be seen. The spatial co-location rules when the neighborhood size is 0km are presented by graph-based visualization in Figure 4.10 and Figure 4.11. In Figure 4.10 (a), the genus pairs are the vertices. there were only 3 genus pairs inside which indicated that Syringa and Cornus, Syringa and Forsythia, Syringa and Cercis were 3 involved genus pairs in the rules. So the characteristic of Figure 4.10 (a) was offering the indication for the main kinds of genus involved in the rules. But the detail of each rule is not easily to see in Figure 4.10(a). Figure 4.11 (a) uses item sets composed of multiple pairs as vertices. From different size of neighborhood, the genus pairs, Syringa and Forsythia, Syringa and Cornus, Syringa and Cercis appear frequently, which indicates that these 3

genus pairs tend to co-locate within the same elevation difference very often. And Syringa is related to many other kinds of genus to form the genus pairs, it also show that Syringa distributes widely among other genus.

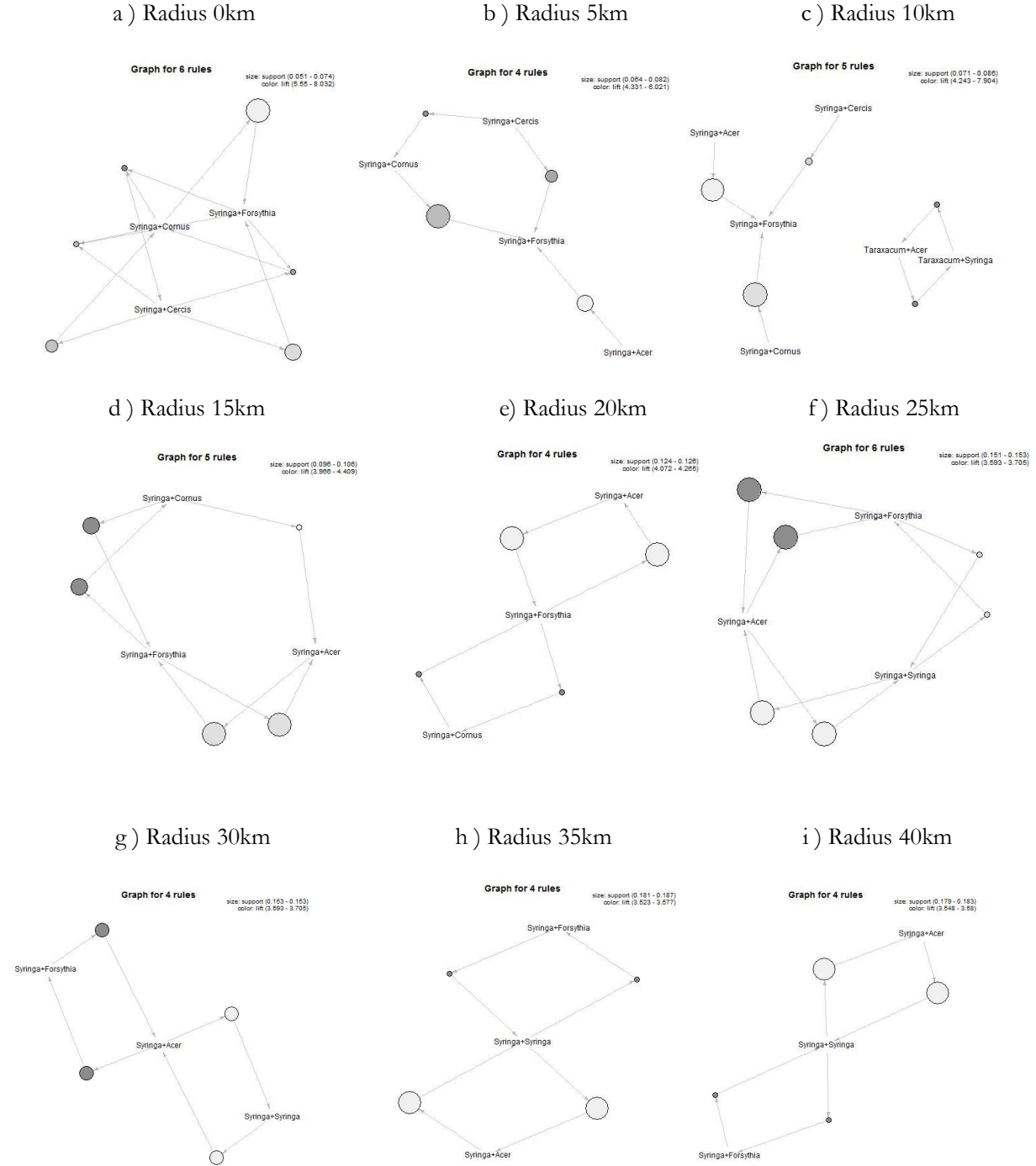


Figure 4.10 Graph-based visualization using items as vertices for spatial co-location rules from distance and elevation

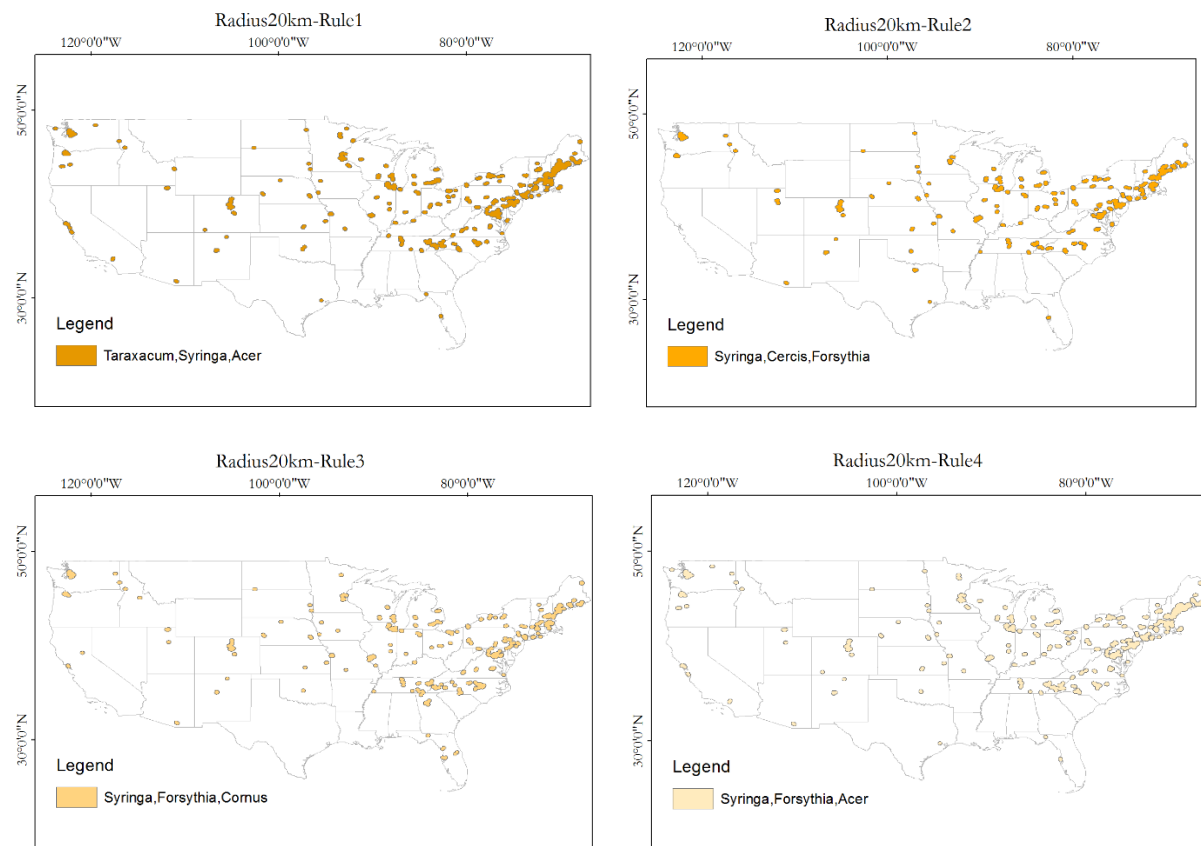



Figure 4.12 Spatial co-location rules map generated by distance and elevation when the distance is 20km

4.3. Spatio-temporal Co-location Mining

4.3.1. Spatio-temporal Co-location Mining Preprocessing

To mine spatio-temporal colocations, the preparation for spatio-temporal transaction was necessary for applying it to Apriori algorithm. The way to generate the transaction data based on distance and the size of neighborhood, and its interval was the same as spatial co-location mining part. In the spatio-temporal co-location mining table, as described in section 3.1, each phenological observation contains the dates of the first and last days when a given phenophase was observed. The sample data from spatio-temporal co-location mining table is shown in Figure 4.13. So if there was time overlapped between two observations, and the overlapped time interval was more than 2 days, we regarded these two kinds of genus as one item. Instead of grouping the timing of observed date when the phenological event was observed directly, for instance, the 1st day to the 20th day was regarded as one time interval, the temporal element was added based on the relative time interval, similar as adding elevation to distance. The transaction data from year 2009 to 2014 were merged into one complete spatio-temporal transaction data. For example, according to the distance among different locations, location (1699918.808,-1739372.705),(1701769.659,-1736698.193) were arranged into one transaction, meanwhile, the time interval for Quercus is from the 58th to 70th day and for Carduelis is from the 62th to 105th day. There was overlapped between these two genus, so Quercus and Carduelis were combined as one item “Carduelis+Quercus” in the transaction data to express that they were in the same temporal neighborhood (Figure 4.14).

Latitude	Longitude	First_yes_doy	Last_yes_doy	Genus
26.20864	-98.108391	81	95	Taraxacum
26.39483	-81.503899	38	122	Cornus
27.72086	-82.433151	31	52	Cornus
27.78029	-82.780289	20	57	Bombus
27.78029	-82.780289	54	54	Cornus
27.78029	-82.780289	58	70	Quercus
27.80074	-82.757439	62	105	Carduelis



X	Y	First_yes_doy	Last_yes_doy	Genus
191261.1729	-2078102.526	81	95	Taraxacum
1850168.851	-1864223.672	38	122	Cornus
1734720.379	-1739002.201	31	52	Cornus
1699918.808	-1739372.705	20	57	Bombus
1699918.808	-1739372.705	54	54	Cornus
1699918.808	-1739372.705	58	70	Quercus
1701769.659	-1736698.193	62	105	Carduelis

Figure 4.13 Coordinates conversion in spatio-temporal co-location mining table

NTID	(X,Y)	Genus
1	(191261.1729,2078102.526)	Taraxacum
2	(1850168.851,-1864223.672)	Cornus
3	(1734720.379,-1739002.201)	Cornus
4	(1699918.808,-1739372.705) (1701769.659,-1736698.193)	Bombus+Cornus, Carduelis+Quercus

Figure 4.14 Spatio-temporal transaction data(sample)

4.3.2. Spatio-temporal Co-location Rules

The spatio-temporal transaction data prepared from different size of neighborhood were the input of Apriori algorithm. Just like the elevation and distance based co-location mining process, finding the best minimum support and minimum confidence helps to mine the convincing rules.

After the combination of genus pairs based on those observation of phenophased time range overlapped, many items (genus pairs) were created by 54 genus. For example, when the distance was 0km, there were 928 number different genus pairs existing in transactional database with density of 0.002. When the distance become

20km, the genus pairs turned to 1561 with density 0.004. It indicated that there were many genus had the time overlapped for the observed phenophase with others. So, the support for the frequent patterns were lower than the support from spatial transaction data. Meanwhile, when the neighborhood size increased, more kinds of genus pairs fall in different distance-based transactions, that increase the frequency for some genus pairs.

Figure 4.15 presents 9 scatter plots of the spatio-temporal co-location rules and its measures, which can give the indication for the minimum threshold selection. Compared with the spatial co-location rules based on distance and elevation, the overall support was lower. The support range on axes seen from the graph were around 0 to 0.07.

To find the appropriate minimum threshold for spatial co-location mining, the scatter plot (Figure 4.9) are presented the rules distribution along with measures. Firstly, although the shading of all the dots were lighter than the dots from distance-based spatial co-location rules scatter plot, the lift range for the rules were much larger than it in distance-based result. But the support range for different size of neighborhood, generally it was quite low. Most were less than 0.2, which shows that the distribution of the co-occurrence for multiple species pairs were really sparsely. Moreover, the darkest dot that indicate the high lift from the graphs exists above 0.5. therefore, the minimum confidence were set as 0.5. The highest lift rules focus on the border of the confidence axis when confidence was over 0.6. Therefore, the minimum confidence set as 0.6.

And the support histogram for all the genus pairs in spatio-temporal transaction data were plotted to find the median support. Table 4.3(a) shows the median support for different size of neighborhood which were the value to initiate the algorithm. After many trials, the best minimum support are shown as Table 4.3(b) to find the co-location rules with high measures.

Table 4.3 Median support derived from support histogram shown in Table (a) and the best minimum support for the spatial co-location rules based on distance shown in Table (b)

a)

Distance/KM	Median support
0	0.020
5	0.022
10	0.032
15	0.04
20	0.055
25	0.060
30	0.070
35	0.070
40	0.08

b)

Distance/KM	min_support
0	0.025
5	0.03
10	0.042
15	0.05
20	0.06
25	0.07
30	0.075
35	0.075
40	0.09

The strongest rule from the spatio-temporal co-location mining when the neighborhood size was 20km were that *Acer*, *Syringa* and *Forsythia* tend to co-locate with lift 9.422. And the genus *Syringa* and *Acer*, *Acer* and *Forsythia* tend to have the similar timing of phenophase and co-locate with each other at the same time. When *Acer* and *Syringa* occurred together in a neighborhood and appear in the similar time, the possibility that *Acer* and *Forsythia* would co-locate in space and time was 64%.

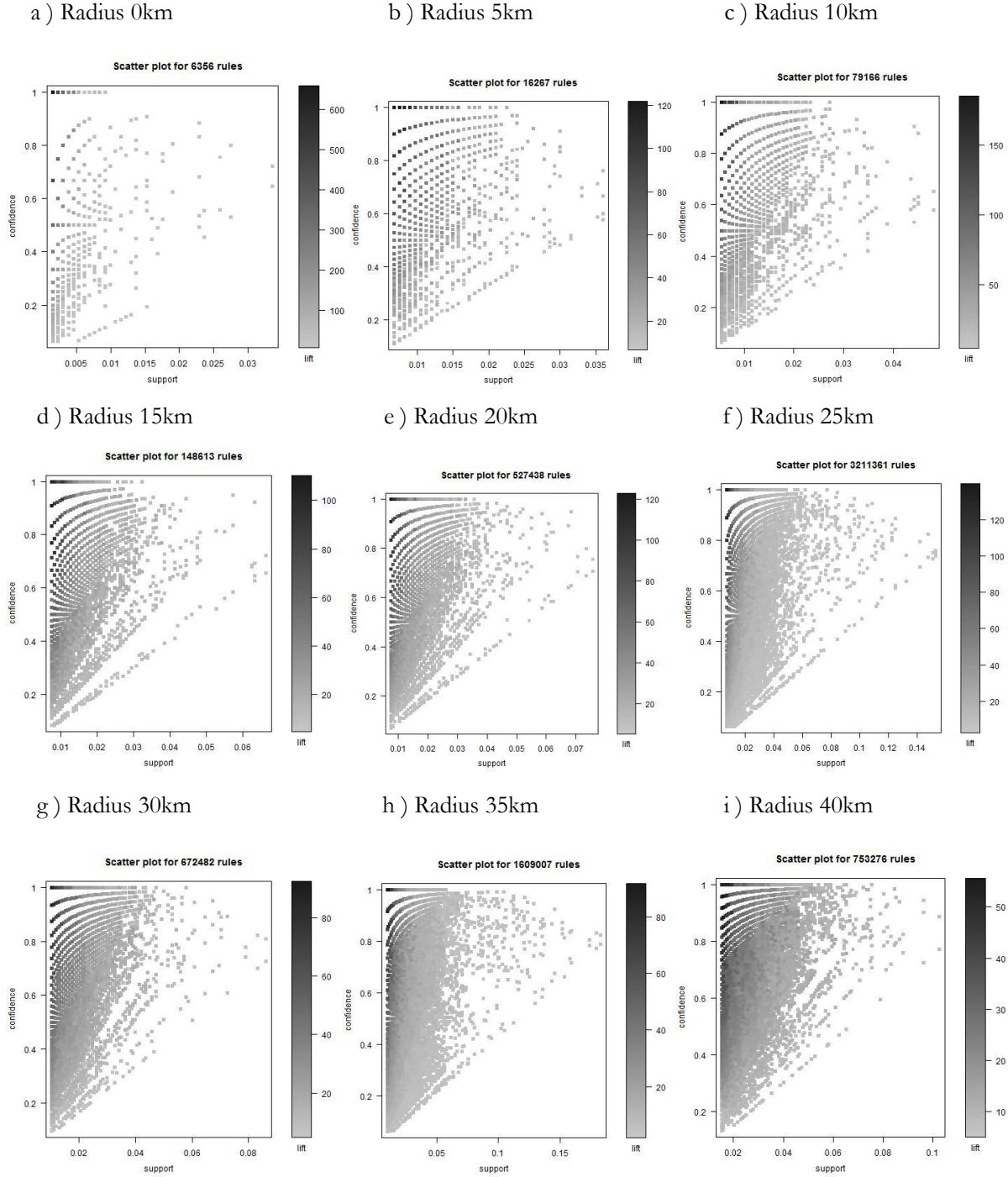
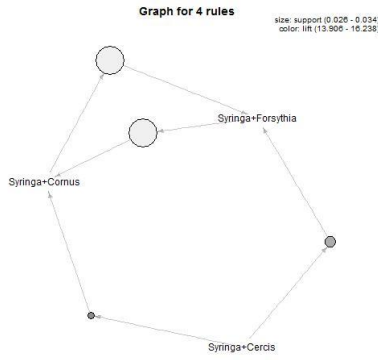


Figure 4.15 Scatter plot of spatial and temporal co-location rules for different size of neighborhood

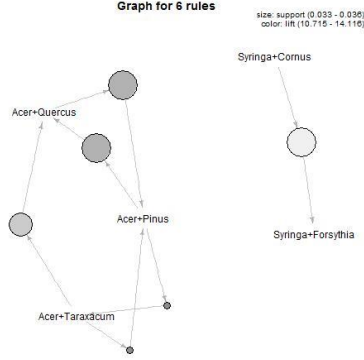
To make the spatio-temporal co-location rules in the format of text visualized, the graph-based visualization were applied to the mined rules. Figure 4.16 and Figure 4.17 presents two-way visualization for the spatio-temporal co-location rules from different neighbourhood. There were two cluttered sets seen in the graphs indicating that part of mined rules were mainly developed among *Acer*, *Forsythia* and *Syringa*, and part of rules were derived from *Syringa*, *Forsythia* and *Cornus* in Figure 4.16(e) when the neighborhood size is 20km. Although *Syringa* and *Forsythia* tend to occur frequent in genus pairs, but the difference from spatial co-

location rules was that the phenophases for Syringa, Forsythia in different pairs were different. As Figure 4.17(e) shown, the co-location relationship in space and time among Acer, Forsythia is the strongest because of the darkest circles in between. And the Syringa, Cornus and Acer related rules are with high support. But in that graph, although the measures of the rules can be seen clearly among genus pairs, but the number of rules can not be seen easily. In Figure 4.17(e), there are 4 co-location rules presented. The duplicated relationship among genus pairs are presented by single line.

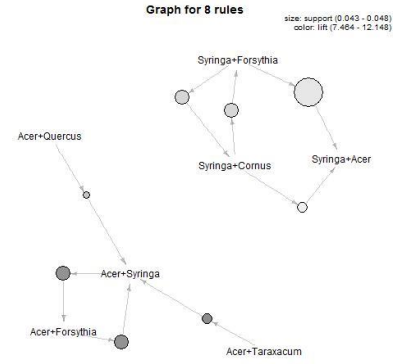
a) Radius 0km



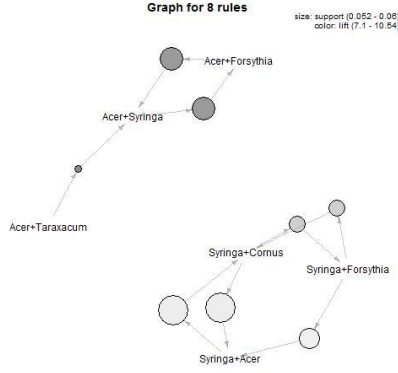
b) Radius 5km



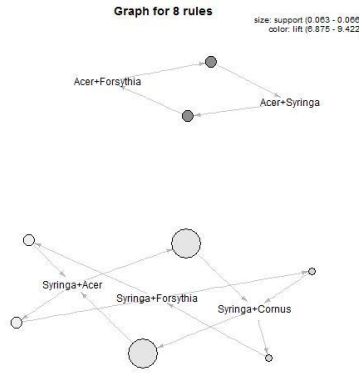
c) Radius 10km



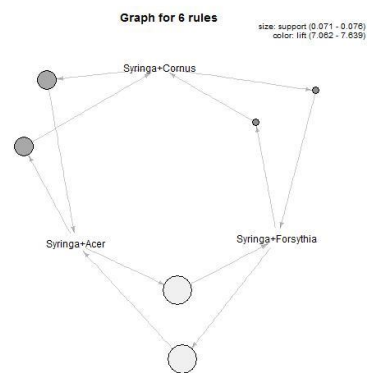
d) Radius 15km



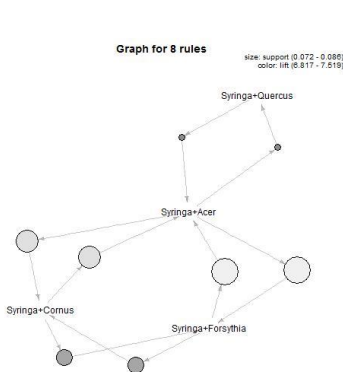
e) Radius 20km



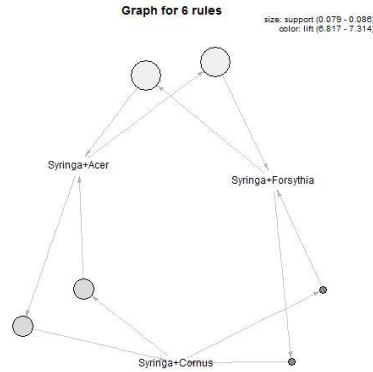
f) Radius 25km



g) Radius 30km



h) Radius 35km



i) Radius 40km

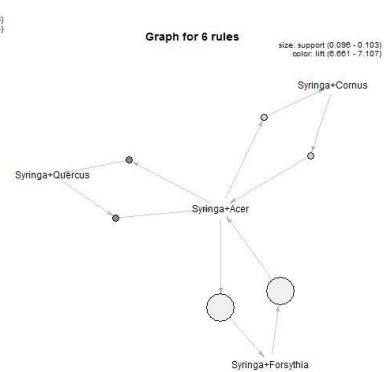
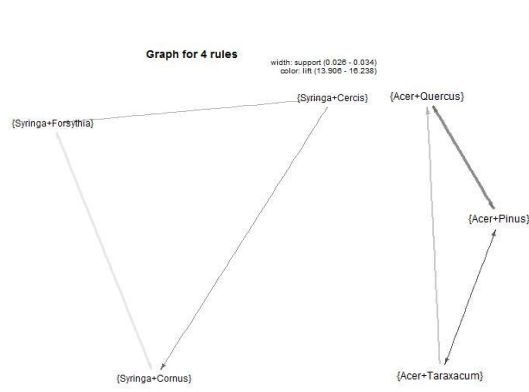
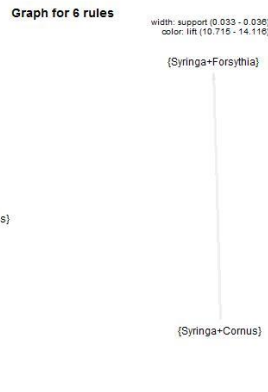


Figure 4.16 Graph-based visualization using items as vertices for spatial co-location rules from distance and elevation

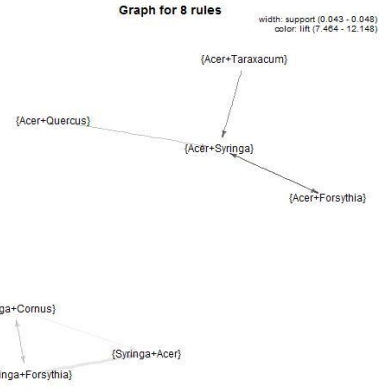
a) Radius 0km



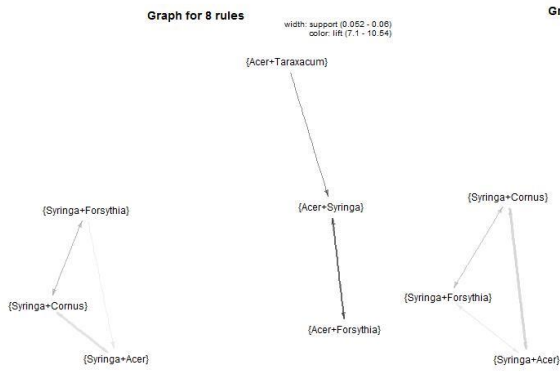
b) Radius 5km



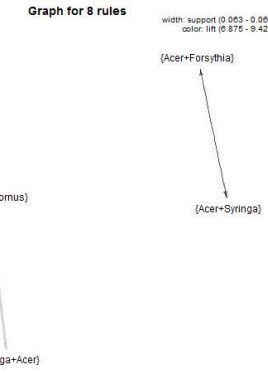
c) Radius 10km



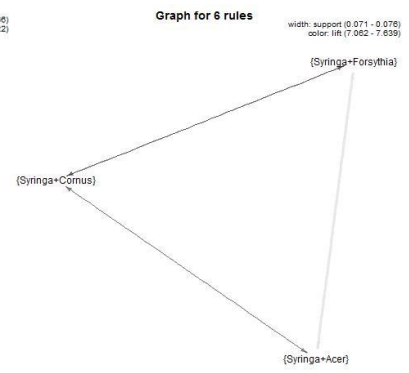
d) Radius 15km



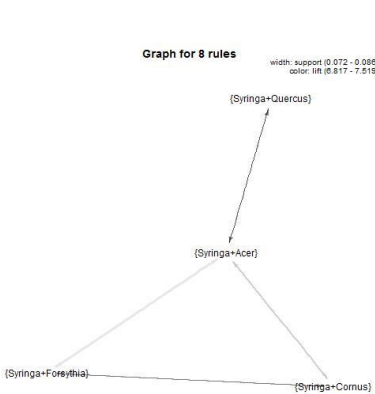
e) Radius 20km



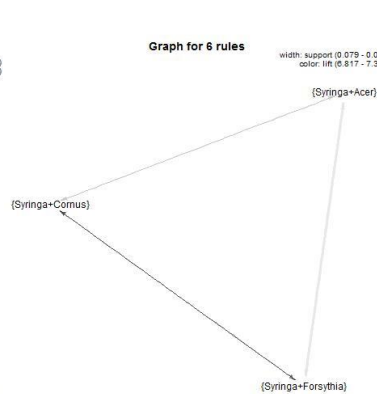
f) Radius 25km



g) Radius 30km



h) Radius 35km



i) Radius 40km

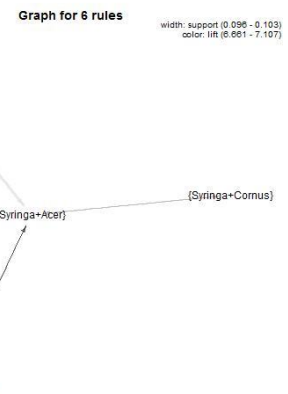


Figure 4.17 Graph-based visualization using itemsets as vertices for spatial co-location rules from distance and elevation

Taking the result from the spatio-temporal colocation rules as an example, when the neighborhood size was 20km, we found that Acer and Syringa, Syringa and Forsythia, Syringa and Cornus, Acer and Syringa, and Forsythia and Acer were the pairs having the similar timings for certain phenophases. To get the information for the phenophase of each genus pairs, the phenophase can be extracted from the spatio-temporal transaction data based on the combination of the genus. In Figure 4.18, the spatial element of the spatio-temporal colocation rules are expressed in the map and its corresponding phenophase for genus and the time range for the phenophase co-occurrence are shown in Figure 4.19.

The phenophase for genus were extracted from the genus pairs. Figure 4.19 illustrates the genus and its phenophase mined from spatio-temporal co-locations rules. And the orange line chart (shown in Appendix 3) shows the timing of phenophase for each involved genus respectively. The blue line chart illustrates the time range for the co-occurrence of genus' phenophase. The falling leaves for Acer last quite long period at the beginning of the year and the end of the year. As investigated, because of the aggregation of time range under one genus type, the time range from the 15th to 92nd is observed from red maple and sugar maple. Other Acer, like silver maple, big leaf maple, fall leaves between July and September. The Syringa's full flowering concentrates on April and June. But the time range of two observations from Syringa's full flowering is from 211th to 250th which is not usual, so these two observations were removed, and it may be the cause for why spring and fall phenological event was mined spatio-temporal co-located. Forsythia's open flowering and Syringa's full flowering are found to happen at the similar time range. As we can see, the timing for Forsythia's opening flowers concentrates on March to June, and September to the end of the year. The time range for opening flowering at the end of the year is caused by several locations in Pennsylvania. The DOY of first and last yeses is from 240 to 344 which makes the time range covers wider in the figure. So during the 92nd to 176th of the year, when Forsythia's opening flowers, Syringa tend to full flowering. Cornus's ripe fruiting and Syringa's open flowers were found to happen together. The duration for Cornus's ripe fruits is from July to October. And the duration for Syringa's open flowers covers relatively wide range. Because several observations from Red Rothomagensis lilac in New Mexico have time range from 85th to 276th of the year, making the Syringa's open flowers a long period and overlapped with fall activity for Cornus. And the time range for Acer's increasing leaf size is really long period. That's because, the observed date for its increasing leaf size of striped maple, and several locations for red maple are among 109 to 325, makes the whole time range for Acer quite wider. It turns that Forsythia's fruiting and Acer's increasing leaf size have overlapping in time.

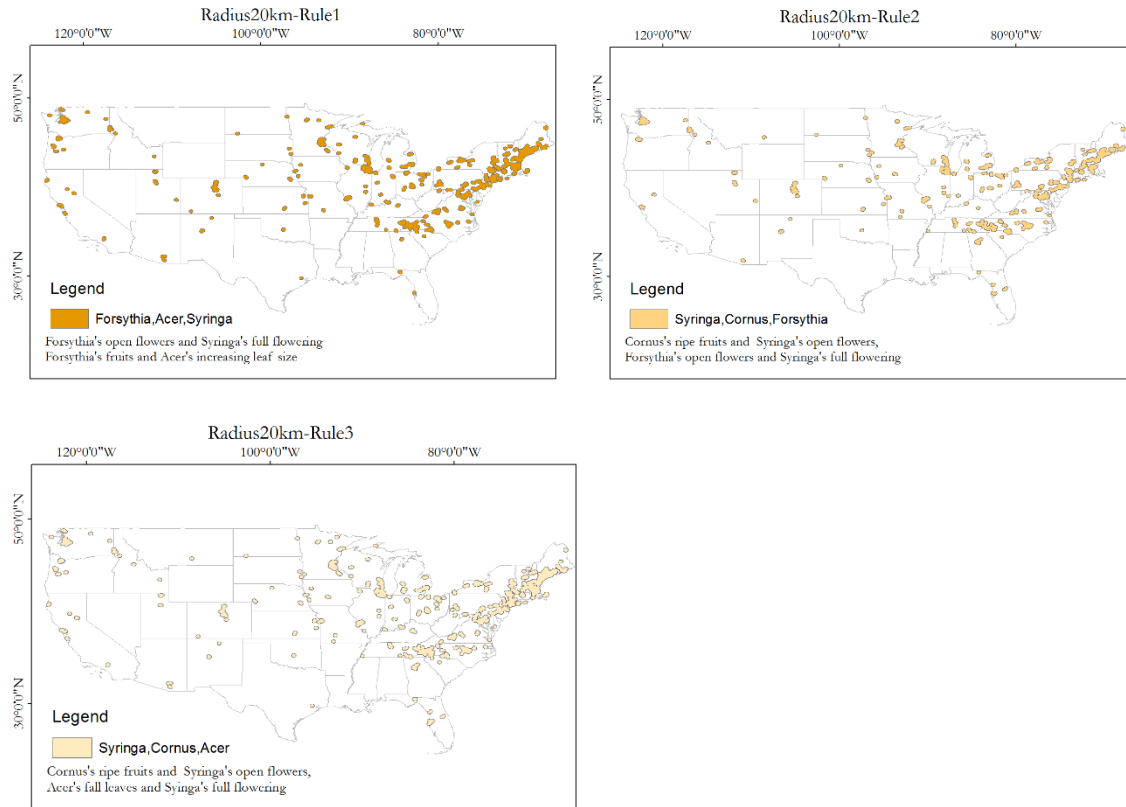


Figure 4.18 Spatio-temporal co-location rules maps when the neighborhood size was 20km

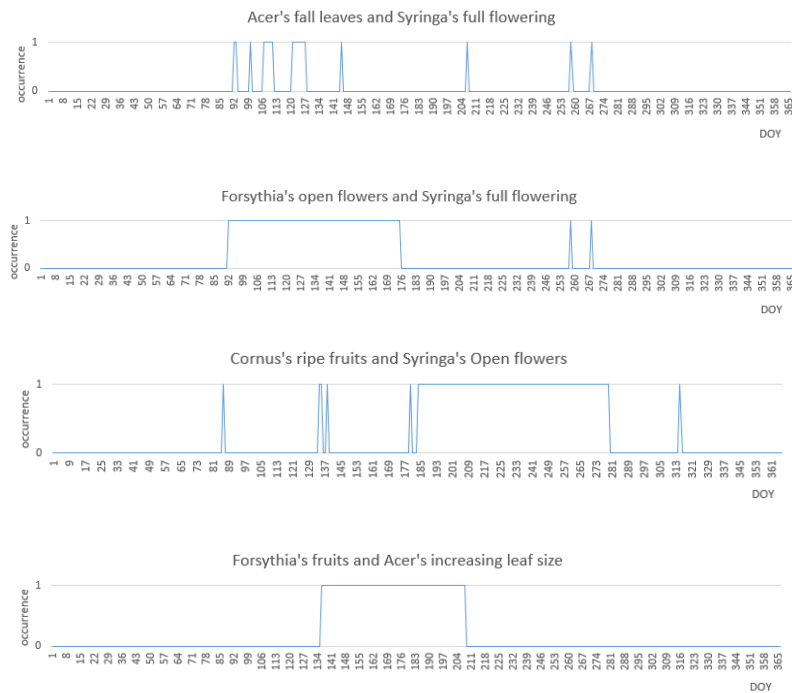


Figure 4.19 Genus pairs phenophase time range for spatial-temporal co-location rules

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

In this MSc thesis I mined spatial and spatio-temporal co-location rules from volunteered phenological observations by using the Apriori algorithm. After that, the rules were visualized in graphs and maps. In this section, we provide answers to each research questions formulated in section 1.

1. Spatial co-locations

1.1. How to prepare the transaction data in space?

Due to the large amount of volunteered phenological observations, it is hard to extract the co-locations directly from phenological database directly. To mine the co-location relationship in large spatial databases, the observations need to be transformed into transactional data. In this research, spatial transaction data were generated in two different ways: 1/ we regarded each unique location in the database as a transaction and the genus appeared in that location were stored to form a transaction database; 2/ Each record in transaction database is one transaction representing an unique location. The transactions are grouped to form a new transaction based on the neighbourhood, of which the radius can be defined by the distance or elevation difference and the genus locate in that specific neighbourhood are stored in the new transaction, which compose the spatial transaction data.

1.2. How to define the neighbourhood in space?

In this research, the neighbourhoods are defined in 2 ways: 1/ the way to group the transactions was based on using distance among locations as the radius of the neighbourhood, which was used to find the genus that co-located nearby; 2/ another way was defined the spatial transaction data based on both distance and elevation. Apart from the grouping based on distance, the process of combing the genus that occur within elevation difference as genus pairs was made. Defining appropriate neighborhood depends on the research purpose and the data. Apart from define the spatial neighborhood by using distance, the elevation can also be used. When the elevation difference of the study area is very small, species distribution may only have little differences in the height, in that case, the elevation may not be a good element to define the neighborhood in space. In general, continuous space in real world is hard to quantify to build the neighbourhood but we only concerns where events happen, like the occurrence in certain neighbourhood in this research.

1.3. How to define the size of the neighbourhood of species to fix the co-location mining area?

In my study, sequent neighbourhood based on 5km distance interval. The distance from 0 to 40km was tested. And the elevation difference was set as 100 meters. The size of neighbourhood using distance was defined by testing many times by checking the mined rules. If the genus that co-locate together tend to be repeating in different neighbourhood size, the distance or the distance interval should be increased. The elevation difference was defined in the same way. Besides, the definition of neighbourhood size also depends on the real data. The data used in my research was from the USA, so co-location mining area with 40km as the radius was not too large while for the country with smaller area, the co-location mining area was too large. The configuration of distance or elevation interval always starts with the test of small interval value. When the interval is too small, it yields the same co-location rules.

1.4. Which species tend to appear near to each other?

Forsythia, Cornus and Syringa were the 3 out of 54 genus that tend to occur nearby when using distance as the radius for neighbourhood, because these 3 genus frequently appear despite the changing of the neighbourhood size. Syringa and Forsythia, Syringa and Cornus, Syringa and Cercis were the mined genus pairs that tend to occur within the co-location mining area and within the same elevation difference. Actually, Syringa, Forsythia are quite common in the dataset but Cornus was not with many observations in the dataset. And many other type of genus like Carya, Fagus which have more observations were not involved in the co-location rules. So the spatial co-location rules do show something different and interesting from the large spatial dataset.

1.5. How to visualize the spatial co-locations?

The visualization of the spatial co-locations was implemented by graph based technique to show the mined rules with its support and lift. Moreover, the involved genus for each rule and its geographic information were reflected in the map by generating the same size of area for the neighbourhood based on the locations. The neighbourhood in distance can be presented in the map by using a buffer for the observations, but when the neighbourhood was too small in the study area, it is hard to present the co-locations by circles.

1.7. How to define the meaningful spatial co-location rules?

The meaningful spatial co-location rules was decided by minimum support, minimum confidence and lift. The rules with high lift tend to be of high quality. The minimum support was configured by analysing the support histogram and the minimum confidence was defined by observing the scatter plot of all the mined rules. For the spatial co-location rules based on distance, the minimum support mainly focus between 0.1 and 0.5 while the confidence was over 0.6 with the lift range from 1 to 2. For the rules based on distance and elevation, the minimum support range was 0.04 and 0.18 while the confidence was over 0.5 with lift around 3.

2. Spatio-temporal co-locations

2.1. How to prepare the transaction data in both space and time?

Spatio-temporal transaction data was firstly by creating the spatial transaction data based on distance, in the same way as spatial co-location mining. After the grouping for transactions, the genus were combined as genus pairs if the relative time interval for a certain phenophase was overlapped. The use of a relative temporal window to find the events that happen at a similar timing avoid the situation when the activities are the constant process.

2.2. How to define the neighbourhood in spatio-temporal way?

The neighbourhood size in space (distance) was defined as same as the process of spatial co-location mining. And the temporal element was defined by the overlapped range between two genus for the phenophase. If the overlapped time range was within 2 days, those two genus was regarded as occur at the same timing of phenophase.

2.3. Which of the spatially co-located species have similar timing of phenophases?

Genus pairs Acer and Syringa, Syringa and Forsythia, Syringa and Cornus tend to have similar timing of phenophase from all size of neighbourhood.

2.4. Which phenophase of the spatially co-located species have when they co-occur with each other?

The phenophase for genus were extracted from the genus pairs involved in mined rules. Compared with all the results mined in spatio-temporal way, Forsythia's open flowering and Syringa's full flowering are found to happen between April and June. Also the fall and spring phenological events were found to occur at the similar timing from the result. Due to the aggregation of the DOYs of first and last yeses from species to genus, it makes the time range for specific phenophase under one genus much longer. Like the red maple and sugar maple fall leaves in early period while other species under Acer starts to fall leaves from September. The observation for red maple and sugar maple makes the period of falling leaves much longer. Besides, the timing of phenophase in different locations varies. The observation information for DOYs of first and last yeses covers a long period, leading to the extension for other species's time range under one genus. And the time range for phenophase in different area may differ and affect the general duaration for the genus. To conclude, the aggregation of DOYs of first and last yeses for different species has strong influnen on the spatio-temporal co-location rules. The aggregation can be improved in 3 ways: 1/the observations that have wide time range for one phenophase should be investigated and checked whether it affects the whole duration. 2/ the sub group can be made under genus. Different species under one genus may have different distribution. The subgrouping can be based on the distribution of various species. Besides, the species may have different phenophase under the same genus, for example, deciduous plants and evergreen plants may be included in one genus type. So the subgrouping can be based on that as well. Before the grouping, the evaluation among the phenophase is really important and necessary. Furthermore, if the number of observations and unique locations are enough to mine the co-locations, the species can be used directly to find the co-location rules without grouping them.

2.5. How to map the spatio-temporal co-locations?

Apart from the same method to visualize the spatial element of the spatio-temporal co-locations, the certain phenophase for involved genus were extracted. The time range for the phenophase occurrence for two genus from spatio-temporal co-location rules was expressed through a line chart.

2.6. How to define the meaningful spatio-temporal co-location rules?

The way to define the meaningful spatio-temporal co-location rules are the same as spatial co-locations by the configurations of parameters, minimum support, minimum confidence and lift. The mined rules in space and time tend to have minimum support range from 0.02 to 0.09, and confidence over 0.6 with lift from 7 to 15. It shows that the possibility of genus's co-occurring in a neighbourhood that derived from spatio-temporal co-location rules is between 0.02 to 0.09. The possibility that when one of the genus occur the possibility that another genus can be found is more than 0.6. Normally, the support range between 0.5% and 30% is reasonable values for many practical domains (Tan, Kumar, & Srivastava, 2002).

5.2. Recommendations

The following recommendations for future work stem from this research:

- The configuration of the parameters is crucial in mining the co-locations. Finding the appropriate minimum support of and the minimum confidence were closely related to the actual data. In my reserach, the support of each genus was not very high because the records in the database are sparsely distributed, especially when mining the spatio-temporal co-locations. Normally, transaction data in the real world tend to have a frequency distribution highly skewed which means almost items occurring with lower support while some of them appear with high frequency the distribution of support. It leads to the situation that some interesting rules are missing because of its low support. An association rule mining algorithm NBMiner, which uses the Negative Binomial Model and mine the co-locations

by an estimated support instead of a user-defined support (Pei, Han, & Mao, 2000).

- Mining spatial and spatio-temporal co-location rules supports the further predications to complete the historic records, due to the incompleteness of phenological database across the years. Each valuable rule that involved certain genus was derived from a certain size of neighbourhood. Therefore, when one genus' occurrence and records were known in the database, another genus from the rule that miss the locations information or the phenophase description can be fixed the location within the neighbourhood or completed the phenophase description.
- In my research, we focus on finding which genus tend to occur next to each other in space and time, which can be regarded as the positive associations between items. However, finding the relationship that when genus A appear, genus B will not occur nearby also can reveal the interaction among species. This kind of association are regarded as negative rules (Ramasubbareddy, Govardhan, & Ramamohanreddy, 2010). In the future, the negative association among species can also be mined to provide valuable information.

LIST OF REFERENCES

- Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases*, 1215, 487–499. <http://doi.org/10.1.1.40.6757>
- Aubrecht, C., Ungar, J., & Freire, S. (2011). Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population, (October), 11–13.
- Batarseh, S. (2014). *Mining Spatio-temporal Datasets Collected by Volunteers : A Case Study Based on US Lilac Data*. University of Twente Faculty of Geo-Information Science and Earth Observation(ITC). Retrieved from <http://ezproxy.utwente.nl:3111/papers/2014/msc/gfm/batarseh.pdf>
- Bayardo, R. J., & Agrawal, R. (1999). Mining the most interesting rules. *ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 145–154. <http://doi.org/10.1145/312129.312219>
- Betancourt, J. L., Schwartz, M. D., & Breshears, D. D. (2007). Evolving Plans for the USA National Phenology Network, 88(19), 211.
- Chuine, I. (2010). Why does phenology drive species distribution? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 365(1555), 3149–3160. <http://doi.org/10.1098/rstb.2010.0142>
- Corey, M. (2013). CHOOSING THE RIGHT MAP PROJECTION. Retrieved from <https://source.opennews.org/en-US/learning/choosing-right-map-projection/#poison>
- De Beurs, K. M., & Henebry, G. M. (2004). Land surface phenology, climatic variation, and institutional change: Analyzing agricultural land cover change in Kazakhstan. *Remote Sensing of Environment*, 89(4), 497–509. <http://doi.org/10.1016/j.rse.2003.11.006>
- De Longueville, B., Luraschi, G., Smits, P., Peedell, S., & De Groeve, T. (2010). Citizens as sensors for natural hazards: A VGI integration workflow. *Geomatica*, 64(1), 41–60.
- Dobson, M. W. (n.d.). VGI as a Compilation Tool for Navigation Map Databases, 307–327. <http://doi.org/10.1007/978-94-007-4587-2>
- Estivill-castro, V., & Murray, A. T. (1994). Discovering associations in spatial data - An efficient medoid based approach. *Science Communication*, 15(4), 457–461. <http://doi.org/10.1177/107554709401500405>
- Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *Geojournal*, 69(4), 211–221. <http://doi.org/10.1007/s10708-007-9111-y>
- Gudmundsson, J., Laube, P., & Wolle, T. (2008). Movement Patterns in Spatio-temporal Data. *ENCYCLOPEDIA OF GIS*, 726–732. <http://doi.org/10.1007/978-0-387-35973-1>
- Hagenauer, J., & Helbich, M. (2013). Hierarchical self-organizing maps for clustering spatiotemporal data. *International Journal of Geographical Information Science*, 27(10), 2026–2042. <http://doi.org/10.1080/13658816.2013.788249>
- Hahsler, M., & Chelluboina, S. (2011). Visualizing Association Rules: Introduction to the R-extension Package arulesViz. *R Project Module*. Retrieved from [http://www.comp.nus.edu.sg/~zhanghao/project/visualization/\[2010\]arulesViz.pdf](http://www.comp.nus.edu.sg/~zhanghao/project/visualization/[2010]arulesViz.pdf)
- Hahsler, M., Grün, B., & Hornik, K. (2005). \pkg{arules} — a Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, 14(15), 1–25. Retrieved from <http://www.jstatsoft.org/counter.php?id=140&url=v14/i15&ct=2;> <http://www.jstatsoft.org/counter.php?id=140&url=v14/i15/v14i15.pdf&ct=1>
- Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 15(1), 55–86. <http://doi.org/10.1007/s10618-006-0059-1>
- Han, J., & Fu, Y. (1995). Discovery of Multiple-Level Association Rules from Large Databases. *Very Large Data Bases - VLDB*, 420–431.
- Huang, Y., Shekhar, S., & Xiong, H. (2004). Discovering colocation patterns from spatial data sets: a general approach. *Tkde*, 16(12), 1472–1485. <http://doi.org/10.1109/TKDE.2004.90>
- Huyen, M., Menne, B., Behrendt, H., & Bertollini, R. (2003). Phenology and human health: allergic disorders. *Health and Global Environmental Change*, 1, 55.
- Inouye, D. W., Barr, B., Armitage, K. B., & Inouye, B. D. (2000). Climate change is affecting altitudinal migrants and hibernating species. *Proceedings of the National Academy of Sciences*, 97(4), 1630–1633. <http://doi.org/10.1073/pnas.97.4.1630>
- Jochen, H., Ulrich, G., Ntzer, & Gholamreza, N. (2000). Algorithms for association rule mining; a general survey and comparison. *ACM SIGKDD Explorations Newsletter*, 2(1), 58–64. <http://doi.org/http://doi.acm.org/10.1145/360402.360421>

- Kondaveeti, A., Liu, H., Runger, G., & Rowe, J. (2011). Extracting geographic knowledge from sensor intervention data using spatial association rules. *ICSDM 2011 - Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*, 127–130. <http://doi.org/10.1109/ICSDM.2011.5969018>
- Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. *Advances in Spatial Databases*, 47–66. http://doi.org/10.1007/3-540-60159-7_4
- Kotsiantis, S., & Kanellopoulos, D. (2006). Association Rules Mining: A Recent Overview. *Science*, 32(1), 71–82.
- Lee, A. J. T., Hong, R.-W., Ko, W.-M., Tsao, W.-K., & Lin, H.-H. (2007). Mining spatial association rules in image databases. *Information Sciences*, 177(7), 1593–1608. <http://doi.org/10.1016/j.ins.2006.09.018>
- Longley, P., & Batty, M. (2003). *Advanced spatial analysis : the CASA book of GIS*. Redlands, Calif: ESRI.
- Mayer, A. (2010). Phenology and Citizen Science. *BioScience*, 60(3), 172–175. <http://doi.org/10.1525/bio.2010.60.3.3>
- Mehdipoor, H., Zurita-Milla, R., Rosemartin, A., Gerst, K. L., & Weltzin, J. F. (2015). Developing a workflow to identify inconsistencies in volunteered geographic information: A phenological case study. *PLoS ONE*, 10(10), 1–14. <http://doi.org/10.1371/journal.pone.0140811>
- Mennis, J., & Liu, J. W. (2005). Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change. *Transactions in GIS*, 9(1), 5–17. <http://doi.org/10.1111/j.1467-9671.2005.00202.x>
- Miller, H. J., & Han, J. (2009a). *Geographic Data Mining and Knowledge Discovery, Second Edition*. Hoboken: CRC Press.
- Miller, H. J., & Han, J. (2009b). *Geographic Data Mining and Knowledge Discovery, Second Edition*. BocaRaton: CRC Press.
- Möller, M., & Gläßer, C. (2011). AUTOMATIC INTERPOLATION OF PHENOLOGICAL PHASES IN GERMANY Department of Remote Sensing and Cartography J . Birger Geoinformation service Birger Hoppbergsblick 12 06118 Halle (Saale), Germany, 4–7.
- Mooney, H., Larigauderie, A., Cesario, M., Elmquist, T., Hoegh-Guldberg, O., Lavorel, S., ... Yahara, T. (2009). Biodiversity, climate change, and ecosystem services. *Current Opinion in Environmental Sustainability*, 1(1), 46–54. <http://doi.org/10.1016/j.cosust.2009.07.006>
- Morimoto, Y. (2001). Mining frequent neighboring class sets in spatial databases. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '01*, 353–358. <http://doi.org/10.1145/502512.502564>
- Muelliganni, C., Janowicz, K., Ye, M., & Lee, W.-C. (2011). Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information. *Spatial Information Theory*, 6899, 350–370. Retrieved from <Go to ISI>://WOS:000307083100019
- New, P. (2005). Implementing a U.S. National Phenology Network, 86(51). <http://doi.org/10.1029/2003GL019232>.
- Pei, J., Han, J., & Mao, R. (2000). An Efficient Algorithm for mining frequent closed itemsets. *Proceedings of ACM-SIGMOD International Workshop on Data Mining and Knowledge Discovery*, 17(5), 652–664.
- Perry, J. N., & Dixon, P. M. (2002). A New Method to Measure Spatial Association for Ecological Count Data. *Ecoscience*, 9(2), 113–141. Retrieved from <http://www.jstor.org/stable/42901477>
- Ramasubbareddy, B., Govardhan, a., & Ramamohanreddy, a. (2010). Mining positive and negative association rules. *Computer Science and Education (ICCSE), 2010 5th International Conference on*, 27–38. <http://doi.org/10.1109/ICCSE.2010.5593755>
- Richardson, A. D., Keenan, T. F., Migliavacca, M., Ryu, Y., Sonnentag, O., & Toomey, M. (2013). Climate change, phenology, and phenological control of vegetation feedbacks to the climate system. *Agricultural and Forest Meteorology*, 169, 156–173. <http://doi.org/10.1016/j.agrformet.2012.09.012>
- Schröter, D., Cramer, W., Leemans, R., Prentice, I. C., Araújo, M. B., Arnell, N. W., ... Zierl, B. (2005). Ecosystem service supply and vulnerability to global change in Europe. *Science (New York, N.Y.)*, 310(5752), 1333–1337. <http://doi.org/10.1126/science.1115233>
- Schwartz, M. D. (1994). Monitoring global change with phenology: The case of the spring green wave. *International Journal of Biometeorology*, 38(1), 18–22. <http://doi.org/10.1007/BF01241799>
- Schwartz, M. D., Betancourt, J. L., & Weltzin, J. F. (2012). From caprio's lilacs to the USA National Phenology Network. *Frontiers in Ecology and the Environment*, 10(6), 324–327. <http://doi.org/10.1890/110281>
- Shekhar, S., & Huang, Y. (2001). Discovering spatial co-location patterns: A summary of results. *Advances in Spatial and Temporal Databases*, 2121, 236–256. http://doi.org/10.1007/3-540-47724-1_13

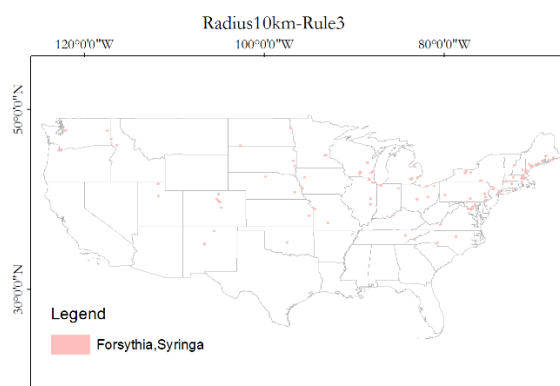
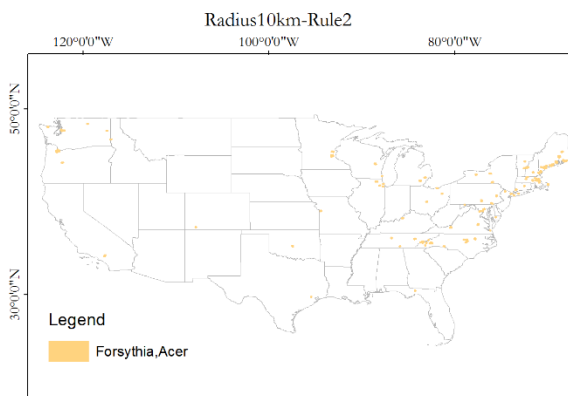
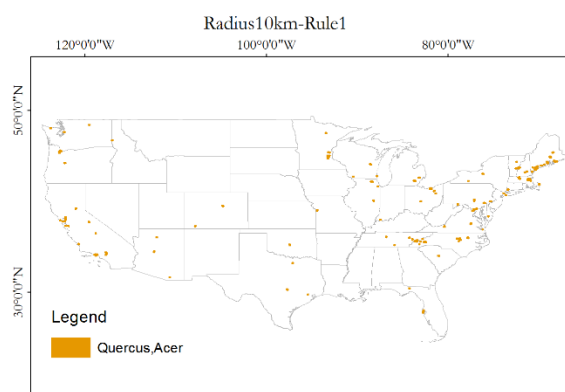
- Stodden, V. (2010). Open science: policy implications for the evolving phenomenon of user-led scientific innovation. *Journal of Science Communication*, 9(1), 1–9. Retrieved from [http://jcom.sissa.it/archive/09/01/Jcom0901\(2010\)A05/Jcom0901\(2010\)A05.pdf](http://jcom.sissa.it/archive/09/01/Jcom0901(2010)A05/Jcom0901(2010)A05.pdf)
- Studer, S., Stöckli, R., Appenzeller, C., & Vidale, P. L. (2007). A comparative study of satellite and ground-based phenology. *International Journal of Biometeorology*, 51(5), 405–414. <http://doi.org/10.1007/s00484-006-0080-5>
- Tan, P.-N., Kumar, V., & Srivastava, J. (2002). Selecting the Right Interestingness Measure for Association Patterns. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '02*, 2, 32–41. <http://doi.org/10.1145/775052.775053>
- USA Naional Phenology Network. (2011). Phenology data overview. Retrieved from <https://www.usanpn.org/>
- Van Vliet, A. J. H., De Groot, R. S., Bellens, Y., Braun, P., Bruegger, R., Bruns, E., ... Sparks, T. (2003). The European Phenology Network. *International Journal of Biometeorology*, 47(4), 202–212. <http://doi.org/10.1007/s00484-003-0174-2>
- Verhein, F., & Chawla, S. (2006). Mining spatio-temporal association rules, sources, sinks, stationary regions and thoroughfares in object mobility databases. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3882 LNCS, 187–201. http://doi.org/10.1007/11733836_15
- Walther, G., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J. C., ... Bairlein, F. (2002). Ecological responses to recent climate change, 389–395.
- Wang, L., Zhou, L., Lu, J., & Yip, J. (2009). An order-clique-based approach for mining maximal co-locations. *Information Sciences*, 179(19), 3370–3382. <http://doi.org/10.1016/j.ins.2009.05.023>
- White, A. B., Kumar, P., & Tcheng, D. (2005). A data mining approach for understanding topographic control on climate-induced inter-annual vegetation variability over the United States. *Remote Sensing of Environment*, 98(1), 1–20. <http://doi.org/10.1016/j.rse.2005.05.017>
- Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., & Yoo, J. S. (2004). A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects. In *Proceedings of the Fourth SIAM International Conference on Data Mining* (pp. 78–89).
- Yanenko, O., & Schlieder, C. (2012). Enhancing the Quality of Volunteered Geographic Information: A Constraint-based Approach. *Bridging the Geographic Information Sciences*, 429–446. <http://doi.org/10.1007/978-3-642-29063-3>
- Ye, Y., & Chiang, C. C. (2006). A parallel apriori algorithm for frequent itemsets mining. *Proceedings - Fourth International Conference on Software Engineering Research, Management and Applications, SERA 2006*, 87–94. <http://doi.org/10.1109/SERA.2006.6>
- Zurita-Milla, R., Van Gijzel, J. A. E., Hamm, N. A. S., Augustijn, P. W. M., & Vrieling, A. (2013). Exploring spatiotemporal phenological patterns and trajectories using self-organizing maps. *IEEE Transactions on Geoscience and Remote Sensing*, 51(4), 1914–1921. <http://doi.org/10.1109/TGRS.2012.2223218>

APPENDIX

1. Spatial co-location rules in maps. Spatial distribution of spatial co-location rules mined different size of neighbourhood. Each map presents each spatial co-location rule. The circle indicates the possible area that the spatial co-location rule occur, and the size of the circles indicate the size of the neighbourhood.

(1) Spatial co-location rules mined based on distance transaction data

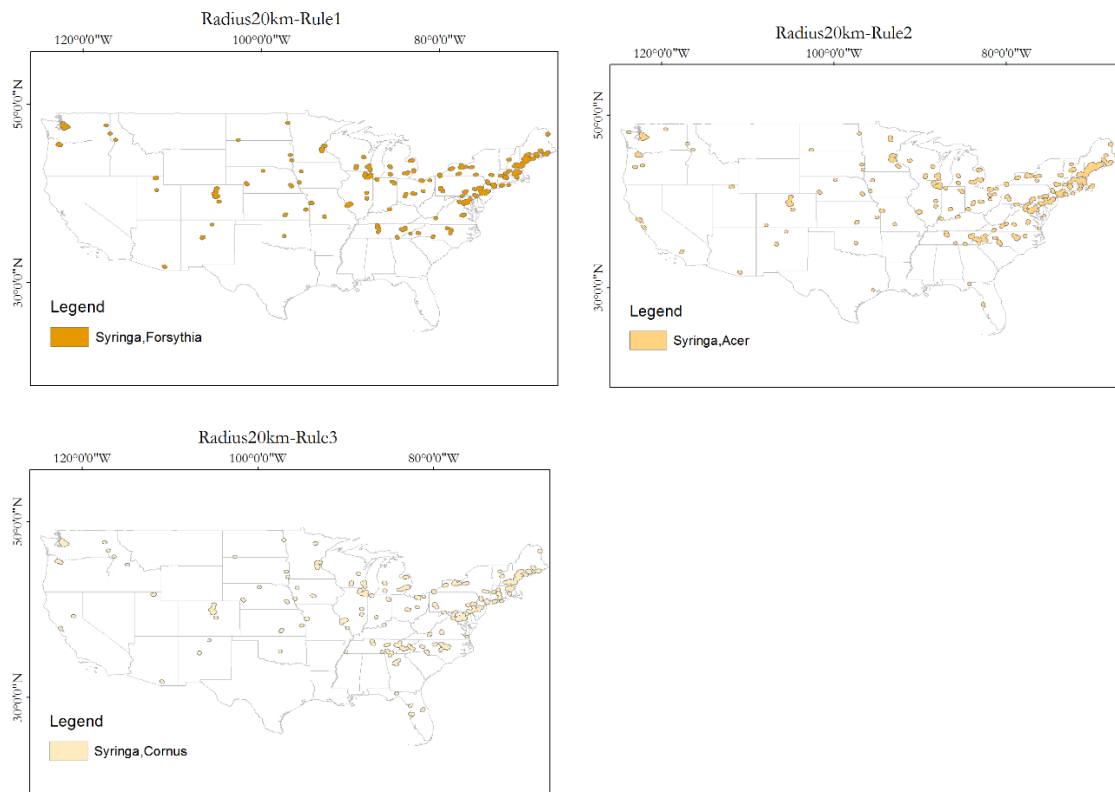
a) Radius 10km



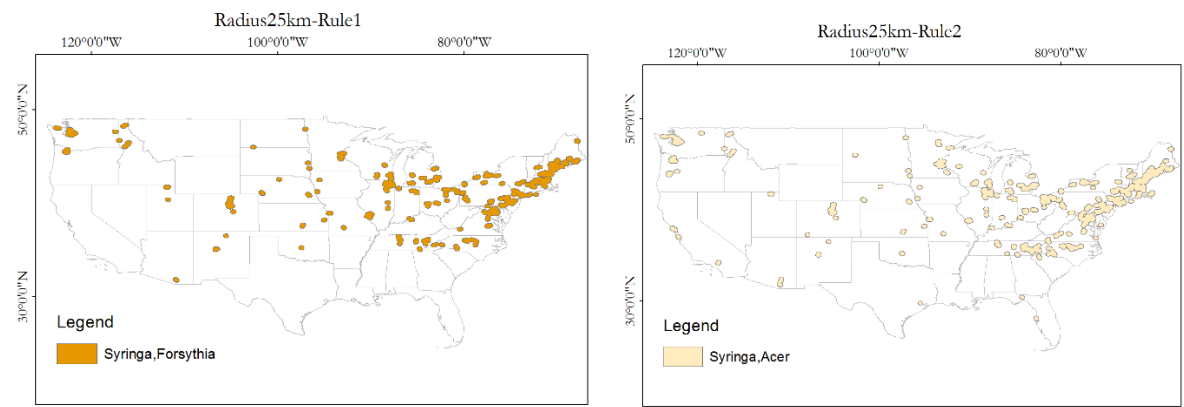
b) Radius 15km



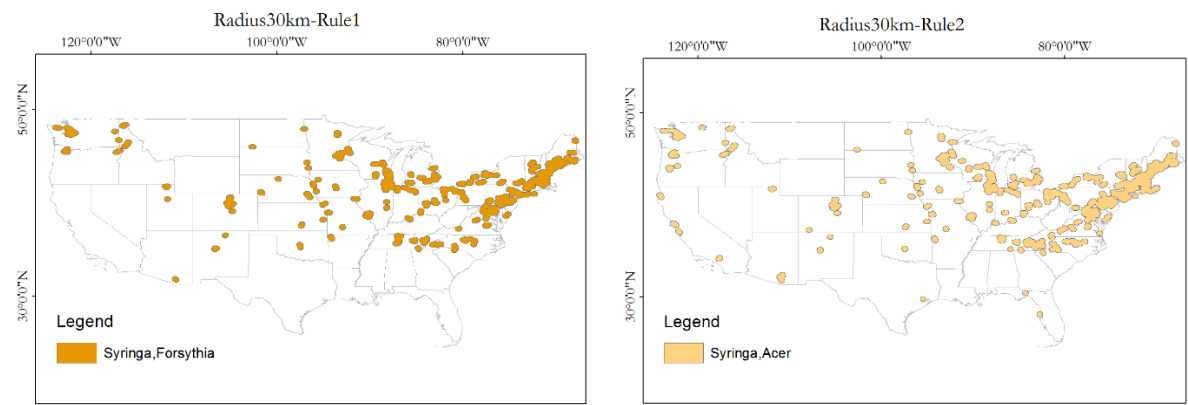
c) Radius 20km



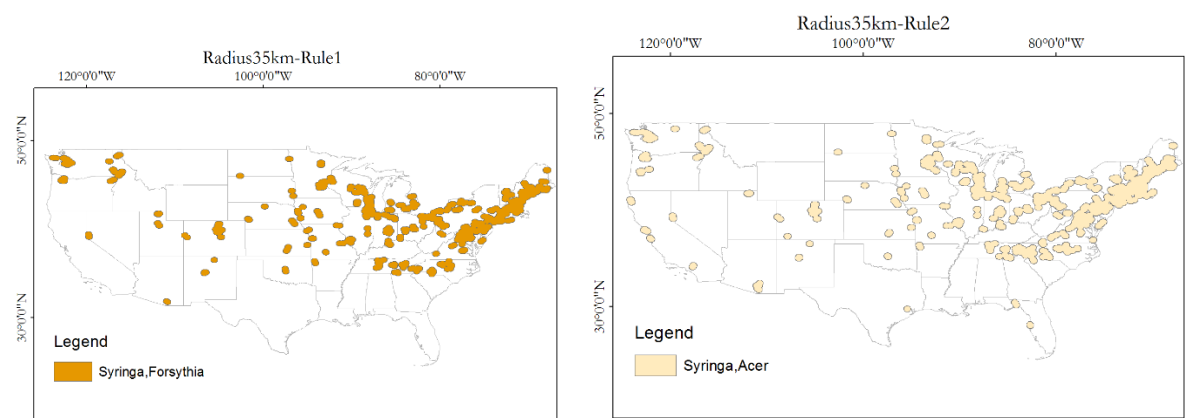
d) Radius 25km



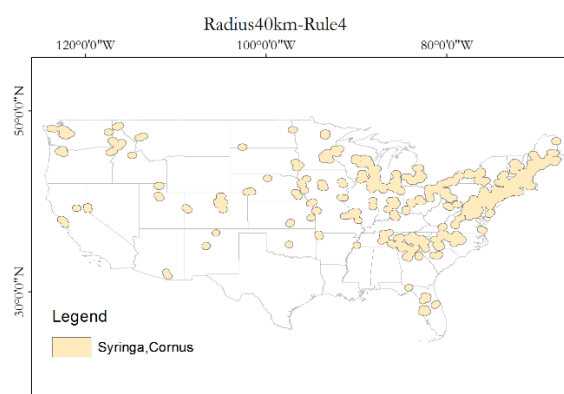
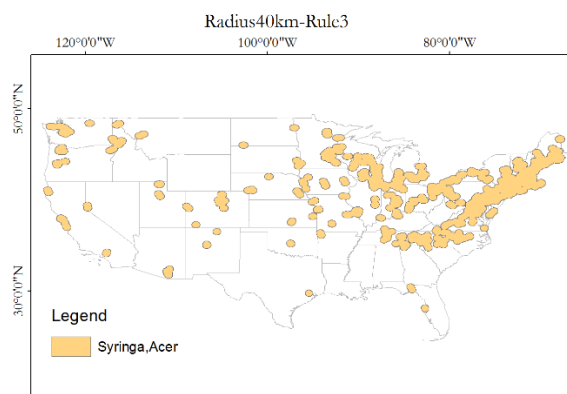
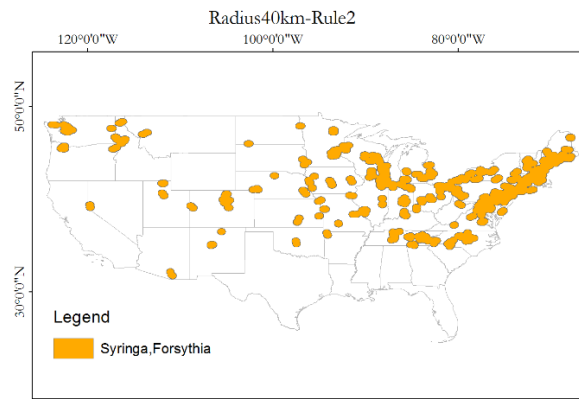
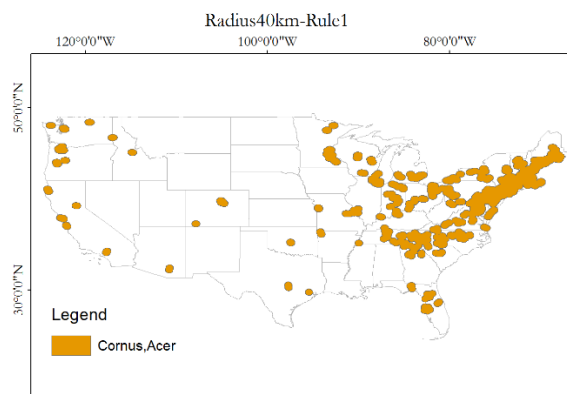
e) Radius 30km



f) Radius 35km

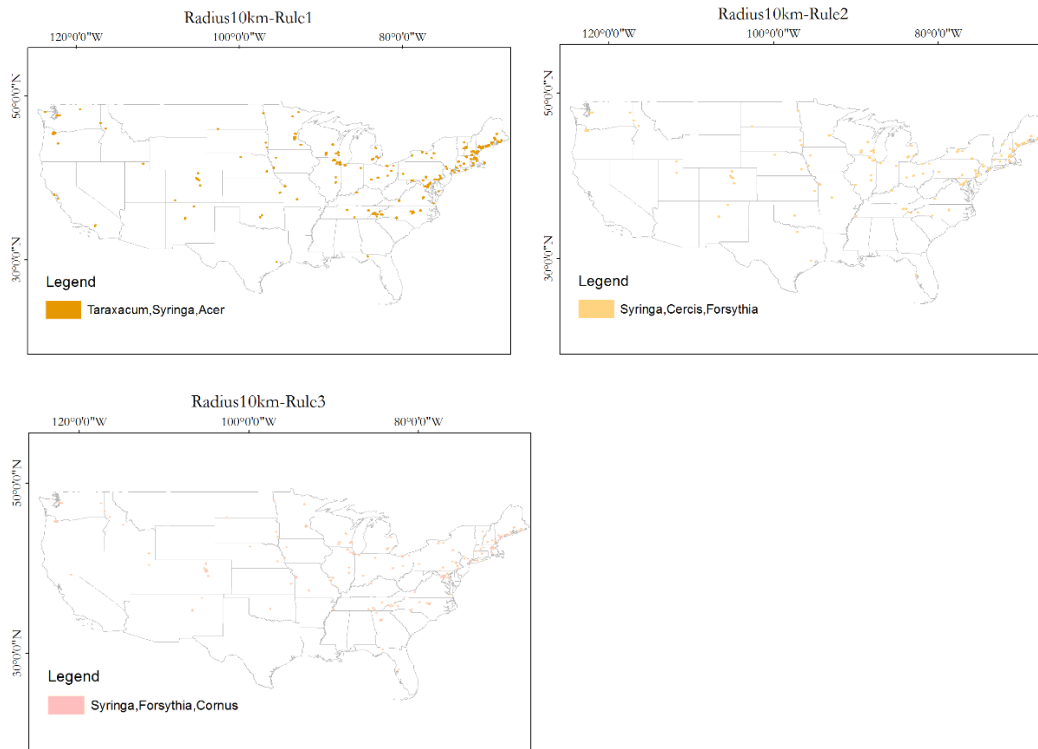


g) Radius 40km

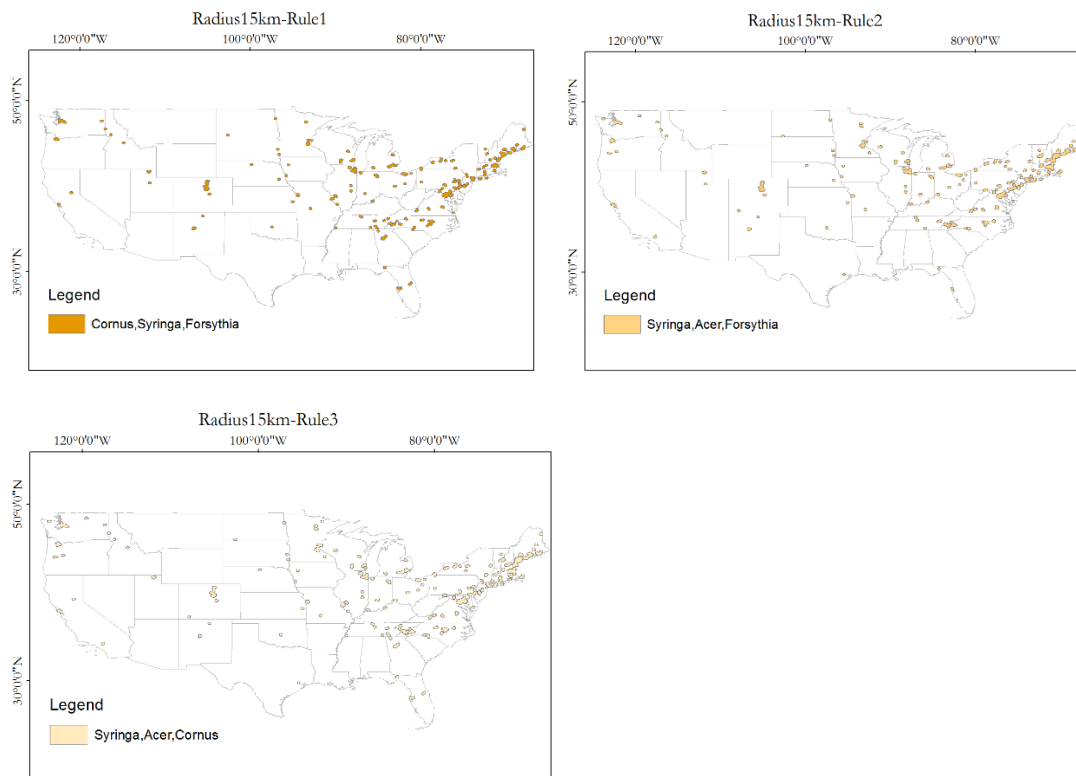


(2) Spatial co-location rules mined based on distance and elevation transaction data

a) Radius 10 km



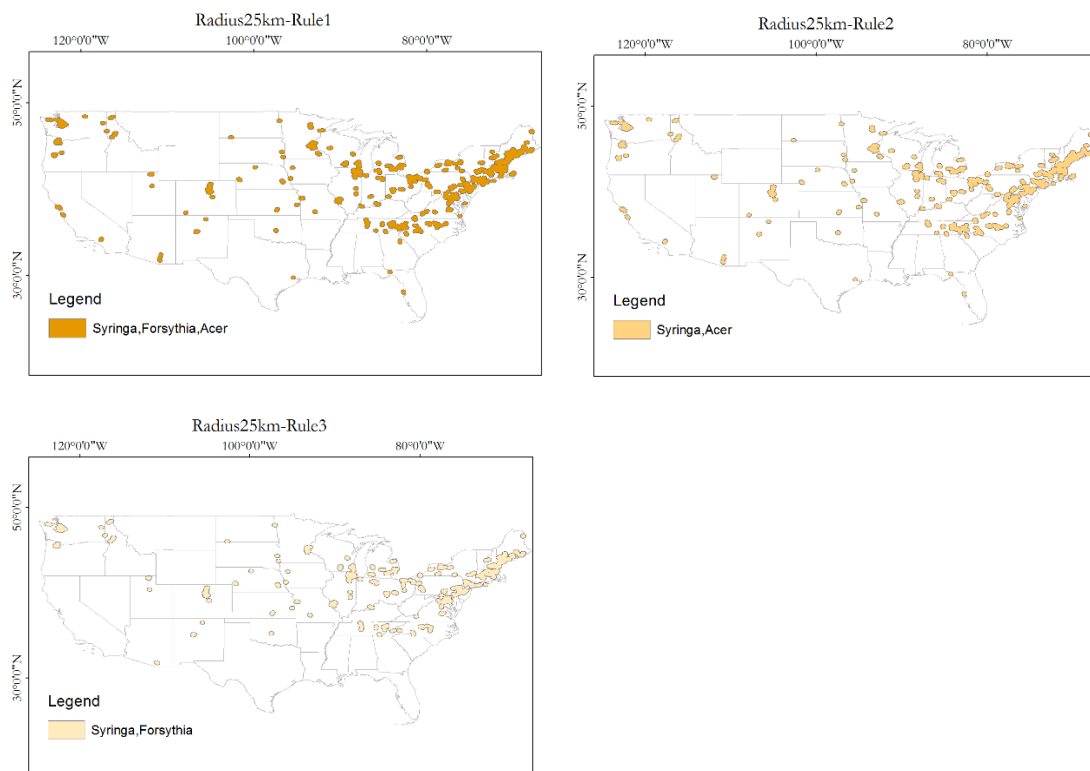
b) Radius 15km



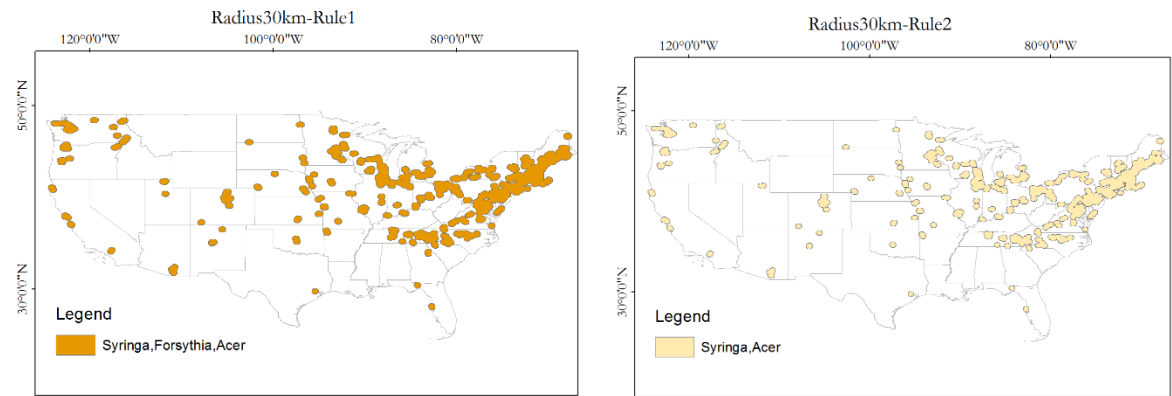
c) Radius 20km



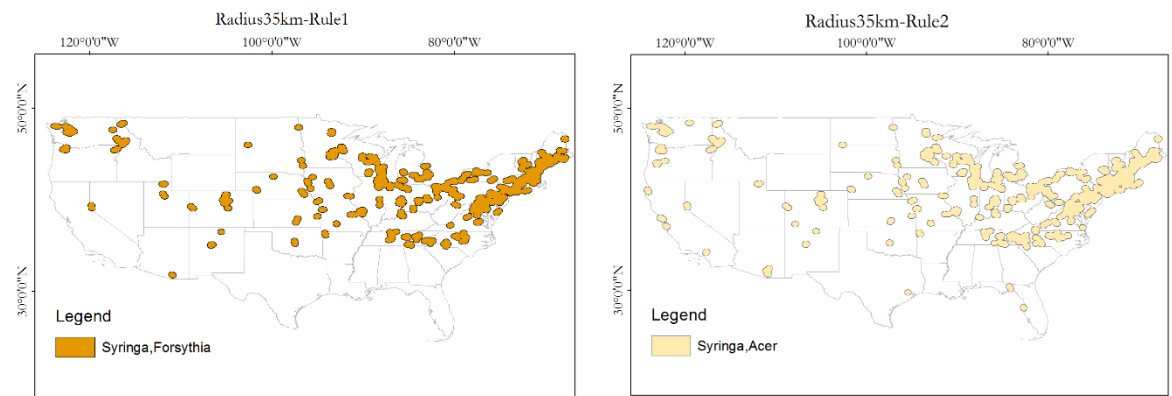
d) Radius 25km



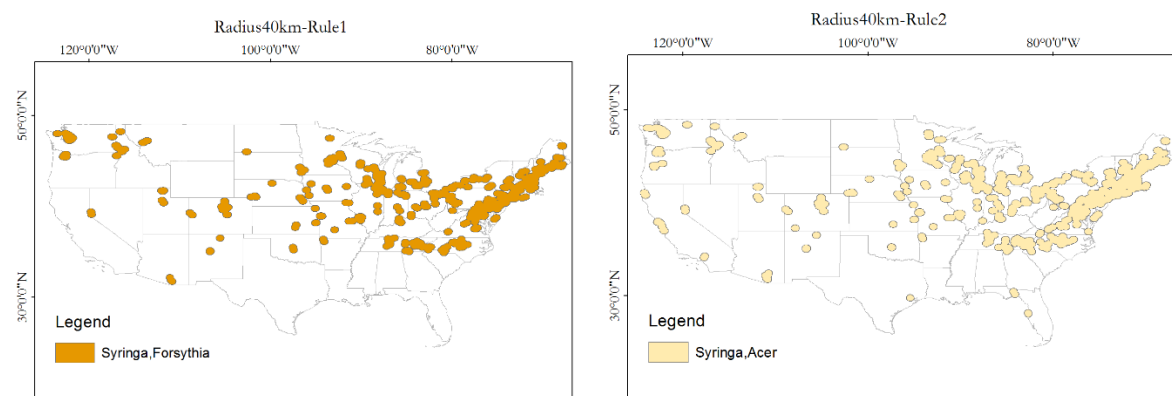
e) Radius 30km



f) Radius 35km



g) Radius 40km

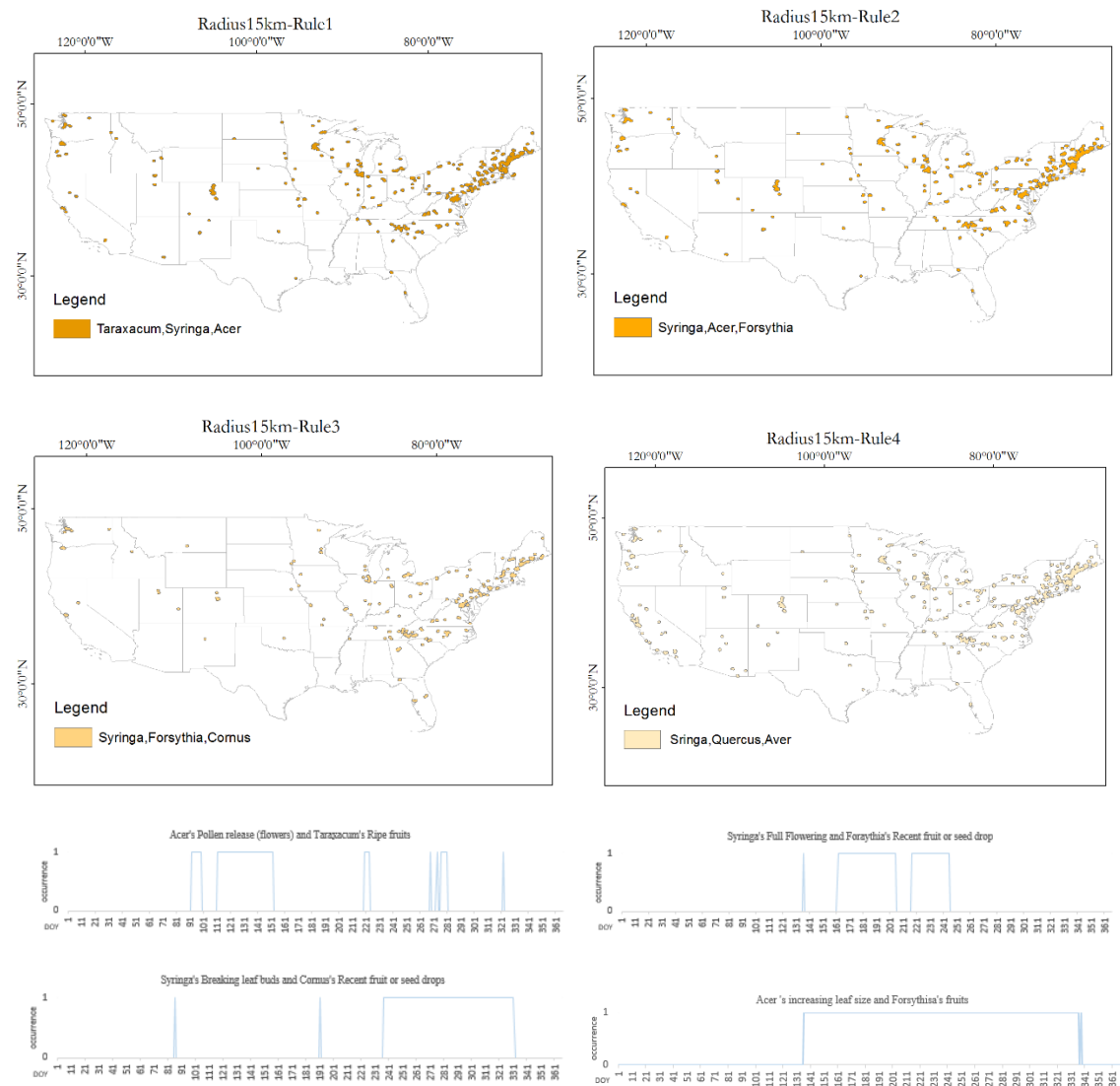


2. Spatio-temporal co-location rules in maps. The spatio-temporal co-location rules are presented in each map and the phenophase time range for the genus in the rules are presented in the graphs

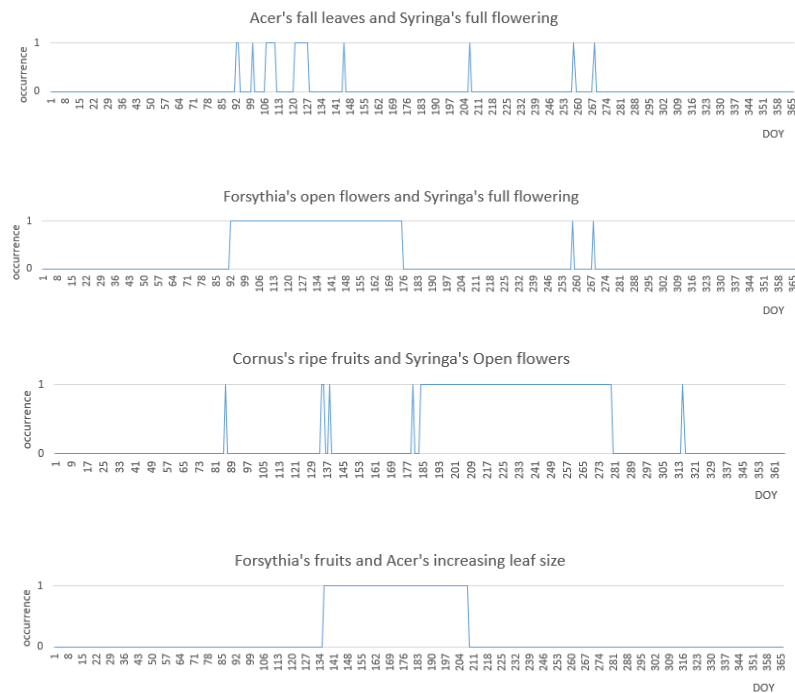
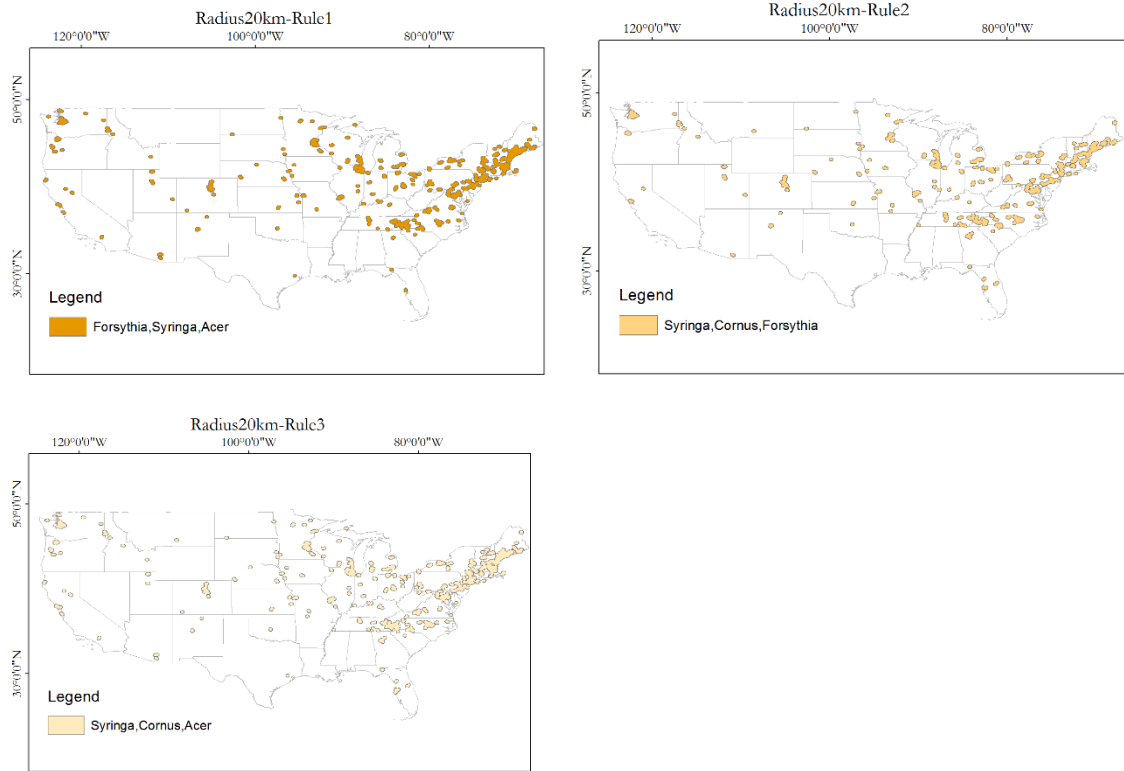
a) Radius 10km



b) Radius 15km



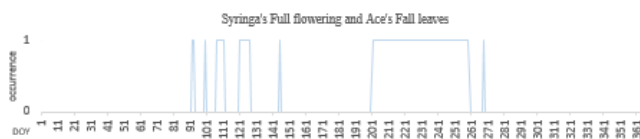
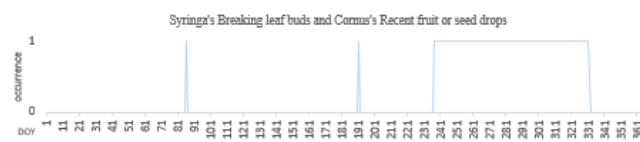
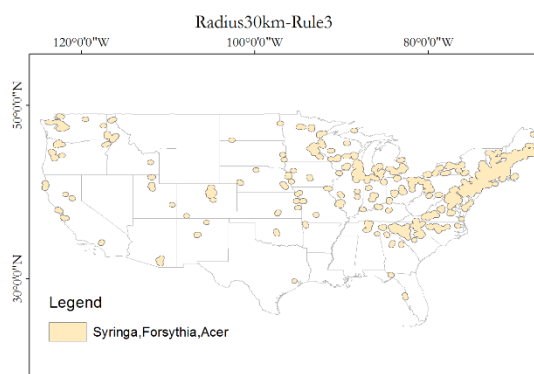
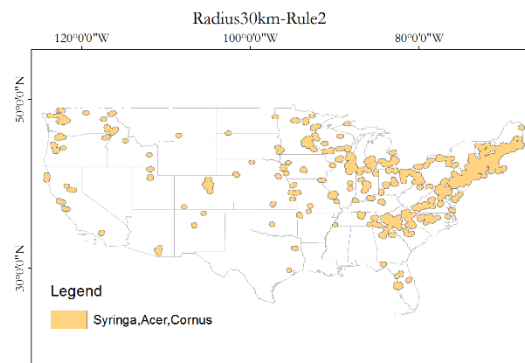
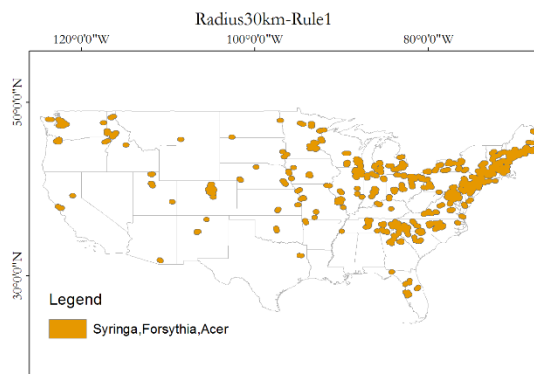
c) Radius 20km



d) Radius 25km



e) Radius 30km



f) Radius 35km



g) Radius 40km



3. The line chart for the timing of phenophase mentioned in Figure 4.19

