Hierarchical Clustering Network Analysis of Ambient Air Monitoring in Alberta - Phases I and 2

Oil Sands Monitoring Program
Technical Report Series 5.0

Canada

Alberta

Hierarchical Clustering Network Analysis of Ambient Air Monitoring in Alberta: Phases 1 and 2

J. Soares[1], P.A. Makar[1], Y. Aklilu[2], A. Akingunola[1]

[1]Air Quality Research Division& Air and Climate Change Policy Branch Environment and Climate Change Canada, Government of Canada

[2]Environmental Monitoring and Science Division Alberta Environment and Parks, Government of Alberta

Recommended Citation:

Soares, J., Makar, P.A., Aklilu, Y. & A Akingunola. 2018. Hierarchical Clustering Network Analysis of Ambient Air Monitoring in Alberta: Phases 1 and 2. Oil Sands Monitoring Program Technical Report Series No. 5.0. 150 p. ISBN: 978-1-4601-4119-9.

ISBN: 978-1-4601-4119-9 (PDF)

October 2018

Table of Contents

Foreword

Since February 2012, the governments of Alberta and Canada have worked in partnership to implement an environmental monitoring program for the oil sands region. In December 2017 both governments renewed their commitment to working together with Indigenous communities in the region by the signing the Alberta-Canada Memorandum of Understanding (MOU) Respecting Environmental Monitoring in the Oil Sands Region. The MOU establishes the foundation for an adaptive and inclusive approach to program implementation ensuring that the program is responsive to emerging priorities, information, knowledge, and input from key stakeholders and Indigenous peoples in the region.

The Oil Sands Monitoring Program is designed to enhance the understanding of the state of the environment and cumulate environmental effects as a result of oil sands development in the region though monitoring and publically reporting on the status and trends of air, water, land and biodiversity. Its vision is to integrate Indigenous knowledge and wisdom with western science to design, interpret, assess, report and govern the program.

Canada and Alberta have provided leadership to strengthen program delivery, and ensure that necessary monitoring and scientific activities meet program commitments and objectives. The oil sands industry provides funding support for the program under the Oil Sands Environmental Regulation (Alberta Regulation 226/2013). Key findings and results from the program inform regional resource management decisions and importantly, are considered as an objective source of scientific interpretation of credible environmental data.

A mandated cornerstone of the program is the public reporting of data, status and trends of environmental impacts caused by development of oil sands resources. The Oil Sands Monitoring Program Technical Report Series provides an objective, and timely, evaluation and interpretation of monitoring data and information collected across environmental media of the program. This includes reporting and evaluation of emission/release sources, fate, effects and transport of contaminants, landscape disturbance and responses across theme areas including atmospheric, aquatic, biotic, wetlands, and community based monitoring.

Executive Summary

The work described herein was carried out under the Network Optimization project of the Oil Sands Monitoring (OSM) program. Network Optimization is the process of examining monitoring data using mathematical analysis techniques, with the purpose of providing information on the data collected at monitoring stations, and on their geographical location.

The driver for the Network Optimization project was the Workshop on Long-Term Air Monitoring Network Optimization, hosted by Alberta Environment and Parks (AEP) in Edmonton, Alberta, in January 2015. The main recommendation resulting from that multi-stakeholder meeting included assessment of potential redundancies in densely clustered areas of ambient air monitoring for both the continuous and passive measurement networks. It was envisioned that the methodology emerging from this assessment could provide scientific advice for optimizing existing monitoring networks, and for designing new networks.

Under the Network Optimization Project, Environment and Climate Change Canada (ECCC) carried out a type of associativity analysis, based on Kolmogorov-Zurbenko (KZ) filtering of the monitoring data and subsequent hierarchical clustering to determine the level of similarity between station records, to provide guidance and advice for the optimization of Alberta monitoring networks. A parallel project is being carried out by AEP, using an additional methodology, removal bias, and will be reported elsewhere.

The purpose of this report is to provide a summary of the ECCC network optimization methodology, and guidance on how its application here to ambient air quality monitoring networks in Alberta may provide insight to aid in air quality network optimization. The scientific advice provided in this report is intended as only one of many inputs to monitoring network optimization.

The report focuses on all the ambient air monitoring networks in Alberta and on two specific oil sands areas: Athabasca and Cold Lake. Alberta's monitoring networks are operated by Airsheds, organizations that monitor and provide public information on air quality, and are identified in the map below (Figure E1). The Airsheds operating in Athabasca and Cold Lake oil sands areas are Wood Buffalo Environmental Association (WBEA) and Lakeland Industrial Community Association (LICA), respectively.

The associativity analysis determines the *level of similarity* between stations' data records based on specific metrics. The analysis is predicated on the concept that stations with highly similar data records over time are potentially redundant (e.g., the most extreme case would be two stations reporting identical data). Here, hierarchical clustering was carried out using two metrics: 1-R, where R is the Pearson's correlation coefficient, and the Euclidean distance. The first metric assessed the similarity in the variation over time of observed concentrations, while the second assessed the similarity in concentration magnitudes. One of the main outcomes of the analysis is an ordering or ranking of stations, according to the degree of similarity of their observation records. Absolute thresholds for redundancy cannot be generated, since the relative rankings depend on the available observation data (number of stations and chemical species observed). The analysis thus does not identify stations which are "redundant" or "not redundant", but rather provides a *relative* ranking of monitoring record similarity, which can in turn be used as one of the inputs for network optimization decision making.

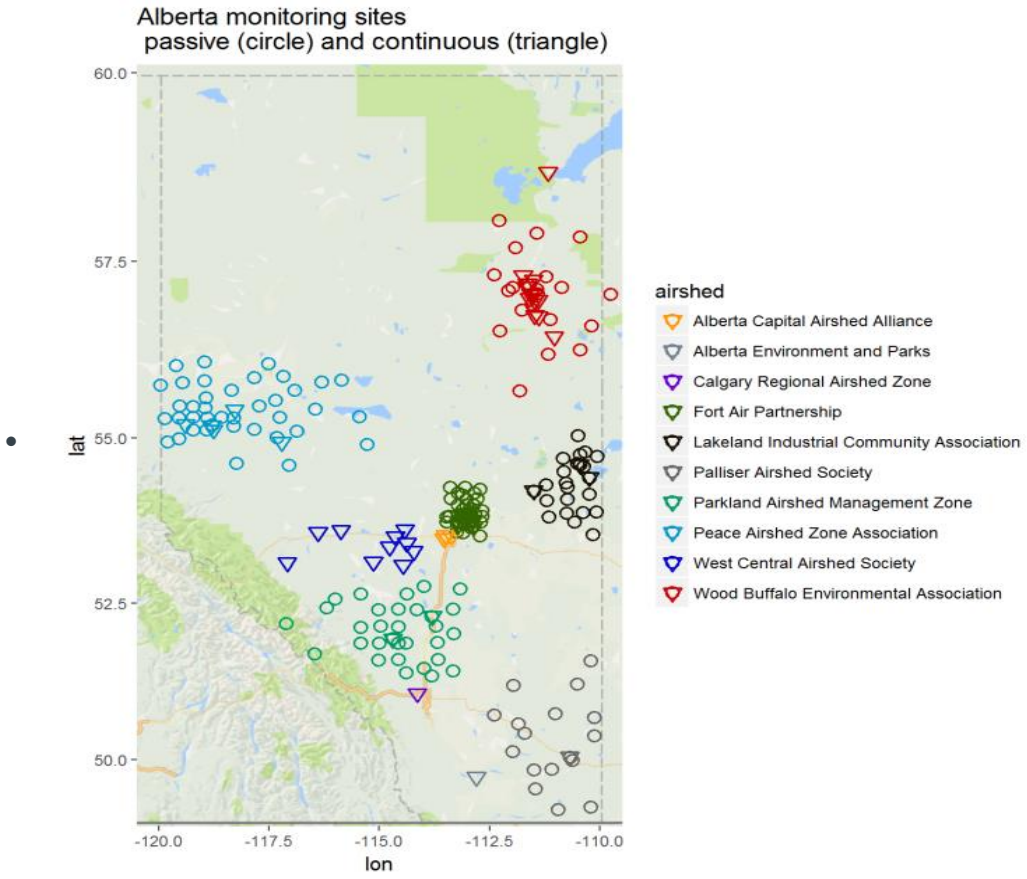


Figure E 1 Air quality monitoring network in the Province of Alberta

The analysis showed that this relative ranking of monitoring station similarity will vary, depending on the chemical, and which of the two metrics was used in the similarity analysis. The information provided by the analysis is thus nuanced. Station records for a given chemical, which are highly similar[1] for both metrics, have a greater degree of relative redundancy than station records which are highly similar for a single metric. Monitoring station locations which are highly similar across both metrics, and across multiple chemicals, have a greater degree of potential redundancy than station locations which are highly similar only for a single chemical. The relative rankings provided in the tables and figures of the report which follows may thus be used in a number of ways, depending on the information considered to be the most important for monitoring network optimization.

As part of the analysis, the records of hourly observations of a suite of chemicals were pre-filtered prior to clustering analysis using an iterative moving average approach (KZ filtering). This additional step was taken to investigate the extent to which the similarity between the station observations was strengthened or weakened as shorter time scales (i.e., high-frequency variation) were removed from the original time

[1] "Highly similar" in this context is with reference to the relative dissimilarity rankings in the tables and figures in the report: i.e. highly similar station records are those which are the *most* similar to those of another station record or cluster.

series. This analysis, and a second analysis in which hourly observations were time averaged to monthly values prior to clustering, showed that much of the similarities between station records were controlled by short term events (concentration changes occurring over hourly or daily time scales) – as opposed to more gradual changes in concentrations. This in turn suggests that observations which consist of long-term averages may be less useful for identifying the unique impact of local emissions sources on a monitoring site compared to hourly observations.

The data from all Alberta Airsheds were collected and quality-control/assured by AEP, then provided to ECCC for subsequent analysis. However, the analysis methodology requires complete station records with relatively few data gaps – some station records could not be analyzed for similarities due to incomplete or missing records, and these stations were identified and the causes of the data gaps discussed.

The analysis methodology was used to provide relative rankings of the continuous (hourly) Alberta monitoring data for the period August 1, 2013 through July 31, 2014, for ozone (O_3), nitrogen monoxide (NO), nitrogen dioxide (NO_2), oxides of nitrogen (NO_X), particulate matter less than 2.5 micrometers in diameter ($PM_{2.5}$), sulphur dioxide (SO_2), non-methane hydrocarbons (NMHC), total hydrocarbons (THC), total reduced sulphur (TRS), and methane (CH_4). KZ-filtering was applied to the hourly data to obtain data series representing daily, weekly and monthly time scales, and all four time scales were analyzed.

The relative rankings of the four most similar and four least similar continuous stations, by station name and chemical, are provided in Table 4.14 and Table 4.15, using the correlation and Euclidean distance metrics respectively. The rankings of all stations analyzed appears in panel (a) of Figures 3.14 to 3.34 (the most similar stations appear at the bottom of these lists). These tables and figures provide guidance on the relative levels of redundancy of the continuous station records, with respect to the two similarity metrics examined here.

These rankings showed a significant variation on the ranking of stations by chemical. For all the chemical species, however, the station records were shown to be more similar for both metrics as shorter time scales were filtered out. This indicates that higher frequency (e.g. hourly) variations in concentration drive most of the differences between the observation data records. This analysis can also be used to show the relative impact of seasonality on station record similarities.

The analysis also showed that hierarchical clustering applied to time-averaged continuous observation records as opposed to time-filtered observation records results in different clusters, and the two operations (time-averaging versus time-filtering) should not be considered as equivalent.

One of the benefits of the analysis approach is its ability to objectively assess the extent of similarities or differences in the records collected via different monitoring methodologies. For this reason, network optimization analysis of the passive monitoring network was carried out using bimonthly passive monitoring data for SO_2 and NO_2, as well as continuous monitoring data for the same chemicals - both passive and continuous data were time-averaged to the same bimonthly level. This examination encompassed the period from February 2009 to December 2015.

This combined analysis identified that the numbers reported from these two sampling methodologies were not similar, at specific levels of the correlation metric, despite similarities in the location, or even collocation, of passive and continuous monitors. Two examples are shown below in Figure E 2.

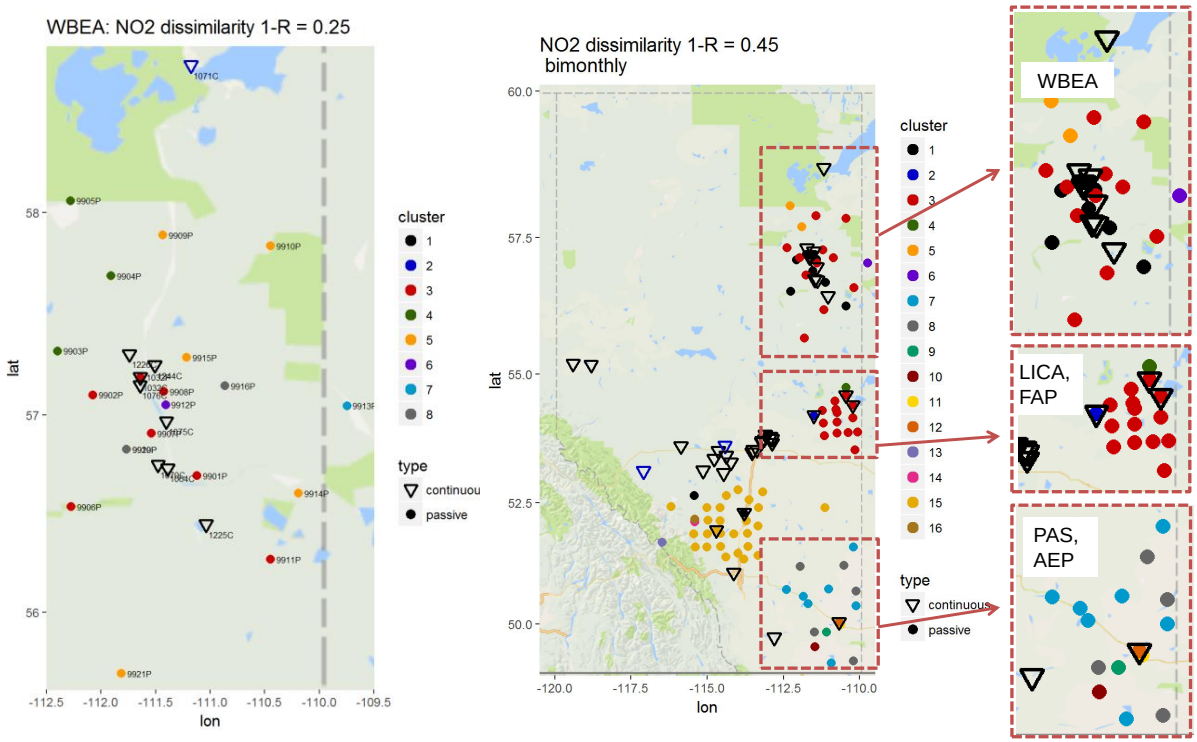


Figure E 2 Clustering of NO_2 monitoring stations for correlation level (left) WBEA monitoring stations, 1-R = 0.25 (R = 0.75), and (right) all Alberta monitoring stations, 1-R = 0.45 (R = 0.55). Clusters are colour-coded within each panel, continuous monitors are shown as inverted triangles and continuous monitoring stations are shown as circles.

In the left panel, continuous NO_2 monitoring stations operated by WBEA are shown as inverted triangles, passive stations as circles, and the colour of the symbols show how different station records were combined by associative analysis into different clusters (each cluster has a different colour). The continuous monitors are all coloured black – that is, they all form a single cluster separate from the surrounding and sometimes collocated passive monitors, at a correlation coefficient of R = 0.75 (1-R=0.25). The right panel shows a similar result for all of the passive and continuous monitors analyzed together in Alberta, at a correlation coefficient level of R=0.55 (1-R =0.45). Collocated passive and continuous monitors once again are not part of common clusters (e.g., at LICA and Palliser Airshed Society (PAS) stations; overlapping black triangles and red dots). The difference in similarity level indicates a larger degree of disparity between the observations reported using the two methodologies at LICA compared to WBEA stations, but both examples show that the passive and continuous methodologies are providing dissimilar observations.

Table 4.1 (NO_2) and Table 4.4 (SO_2) provide both metrics' similarity rankings for bimonthly averaged data for passive and continuous monitors within the WBEA Airshed. Table 4.5 and Table 4.8 are the corresponding tables for the LICA Airshed, and Table 4.11 and Table 4.12 provide the corresponding information for Alberta as a whole. The analysis was shown to be potentially capable of identifying differences in station records associated with type of emissions (stack emissions versus surface sources), with NO_2 showing more disagreement in the order of ranking between the two metrics than SO_2, with the latter having highly correlating stations tending to have a greater discrepancy in their Euclidean distance. The SO_2 analysis of passive monitors also showed a lower degree of similarity (more clusters) between stations at a given level of the similarity metric than the corresponding NO_2 analysis. These findings, and related analyses using the continuous monitoring station records, suggest that the methodology is capable of identifying differences in concentration records relating to the emissions source type, for reasons described in detail in the report.

Section 4 also describes a process by which the results of both metrics may be combined to identify stations with the greatest degree of similarity across both metrics (see the discussion surrounding Table 4.13). The members of each cluster of stations which have been ranked as having relatively high similarity according to correlation may be assessed for the maximum and minimum values of the Euclidean distance. Groups of stations which are relatively highly correlated and have relatively low Euclidean distances may thus be identified; these stations have the greatest degree of potential redundancy from the standpoint of both metrics.

The clustering methodology was found to be sensitive to the precision of the recorded data through three different avenues of investigation. The clustering analysis of SO_2 in particular was found to be strongly impacted by random noise, due to the large number of low concentration data close to the detection limit. O_3 was less impacted, as measured concentrations of O_3 tend to be on the tens of ppbv.

We identified caveats on the accuracy of the observation data, and give recommendations on how the data may be used as an aid in assessing station redundancy in Section 5. Generally we note the following:

(A) The analysis groups stations according to the degree of similarity between stations' data records, but not the cause of that degree of similarity. For example, data records from stations which are separated by large distances, yet are located near emissions sources that happen to have a similar time variation in emissions levels, will be identified as highly similar with respect to the correlation metric. The analysis results should therefore be interpreted with knowledge of local conditions.

(B) There are other constraints associated with monitoring network design, for example geographical factors such as the availability of electrical power and roads, the spatial proximity to highly populated locations or sensitive ecosystems, and the intended purpose of the stations, which are outside of the scope of the current work, yet which are acknowledged here as being important parts of the decision making process.

(C) While passive and continuous monitors were time-averaged to a common bimonthly interval for the purposes of assessing the degree of similarity between the two measurement methodologies, that part of the analysis (Sections 3.1, 3.2, 3.3, 4.1.1, 4.1.2, 4.1.3) was not intended as an assessment of potential

relative redundancies for the continuous monitors—for the latter, the separate analysis of continuous monitoring data (Sections 3.4, 4.1.4) should be examined.

(D) An analysis of the impacts of averaging time on clustering results suggests that the use of observations which comprise long-term averages will reduce the information needed to be able to distinguish records from monitors within an Airshed as being uniquely impacted by sources within that Airshed. Airshed-specific events usually happen on time-scales shorter than monthly averages, for example. The methodology will still correctly identify the relative levels of similarity between monthly data, but the extent to which these similarities are meaningful may be reduced, due to the averaging time associated with the observations. The methodology has the maximum benefit in assessing redundancies when the maximum amount of information is available (i.e., hourly data).

(E) The analysis is limited to the available stations which meet the data completeness criteria and the time period of the data used for analysis. Some stations have been excluded due to the data being insufficiently complete for analysis, and the analysis may be limited by the accuracy (precision) of the methodologies being used for data collection. Stations which were rejected from the analysis due to incomplete data are described in Table 2.2, Table 2.3, Table 2.4, and Table 2.5. We note that the lack of useable data may also be a potential consideration for network optimization.

Despite these caveats, the clustering methodology using hourly data was able to identify groups of stations influenced by common emissions sources (e.g. stations which are influenced by oil sands emissions as opposed to stations located elsewhere), observation records generated using different monitoring methodologies, as well as monitoring station records which were markedly different from all others in the data. The latter may indicate unique recorded events or data inaccuracy; the methodology thus identifies which station records might be worthy of follow-up examination.

Based on the above analyses, we recommend that the assessment of potential redundancies using the tables and figures in this report should be carried out on a “per chemical”, rather than “per station” basis, for stations where more than one chemical species is observed. The clustering analysis of hourly continuous data showed that different chemical species cluster differently, that is, the most similar “stations” for one chemical species may be less similar for other chemical species.

The two metrics may be used separately or together, though we recommend the use of both metrics for assessing potential redundancy whenever possible. The metric chosen for determining redundancies may depend on whether variation in concentration over time or concentration magnitude is considered to be more important with regards to the intent of the monitoring network. However, combinations of the metrics are recommended in assessing potential data record and station redundancies.

Follow-up work to that reported here is taking place at ECCC, and will be reported on at a later date. This work centers on combining output from the air-quality forecast model Global Environmental Multi-scale – Modelling Air-quality and CHemistry (GEM-MACH) with hierarchical clustering, to design air quality monitoring networks which are optimized to reduce similarities between station records. The clusters resulting from the analysis of model-predicted air pollutant time series at observation station locations are being compared to the clusters from the observation data in order to evaluate the model’s ability to mimic observed similarities. The key analysis of this work will be the treatment of all model grid-cells as potential

observation station locations, with the key outcome being maps of optimized monitoring networks to aid in the placement of future air quality monitoring stations. These maps may be combined with other georeferenced data to assist in monitoring network design.

1 Introduction

1.1 Background

The work described herein originated in response to the *Workshop on Long-Term Air Monitoring Network Optimization* in January 2015. Recommendations for short term (high priority) resulting from that multi-stakeholder meeting included assessing redundancies in the densely clustered areas of monitoring using a combination of correlation analysis and/or removal bias with the area served and emissions served information for the continuous measurement network data for SO_2, NO_2, H_2S, THC, and TRS (especially along the Athabasca River Valley and Conklin) and for the passive measurement network data, including industrial sites (this applies to the entire Oil Sands domain). Other actions may be required to address the location of stations. For short to long term actions it was recommended to assess acid deposition and nitrogen deposition monitoring stations and re-design the acid deposition monitoring network, if necessary. There has also been interest expressed by Environment and Climate Change Canada (ECCC) and Alberta Environment and Parks (AEP) towards the development of methodologies which could aid in determining the best possible locations for monitoring network stations.

1.2 Scope

This report will focus on the ECCC project: the use of associativity analysis, specifically hierarchical clustering using metrics of 1-R (*correlation analysis),* and the Euclidean distance, to analyze station data, suggest possible redundancies, and suggest potential "best" locations for future monitoring network stations. The analysis methodology is described in detail in the sections which follow and the Appendix for this report.

The ECCC work has four stages:

(1) Numerical testing of the time-filtering and clustering methodology.

(2) Application of the methodology to AEP monitoring network data.

(3) Application of the methodology to output from the ECCC Global Environmental Multiscale – Modelling Air-quality and CHemistry (GEM-MACH) model, at monitoring network locations.

(4) Application of the methodology to GEM-MACH gridded output.

This report describes the results of the first and second stages of the project – at the time of writing, the third and fourth stages are still underway, and will be the subject of later reporting. This report will include relative rankings of stations based on the degree of similarity of their reported data, as one method of assessing potential redundancies of the existing continuous and passive monitoring network stations, along with caveats regarding the limitations of the analysis, and reporting on issues worth noting which arose during the analysis.

This work is intended to provide scientific advice and analysis to aid in network optimization, but is not intended to be the only means by which network optimization decisions are made. There are other constraints associated with monitoring network design, which are outside of the scope of the current work, yet which are acknowledged here as being part of the decision-making process. These include geographical factors such as the availability of, for example, electrical power and roads, which may limit station locations to sites where these accessibility factors are readily available. Also outside the scope of the present work is the intended purpose of the stations. For example, stations may be required to be placed within a certain distance of emitting facilities due to emissions compliance regulations, as opposed to the extent to which the collected data may be more or less similar to data collected by other monitoring stations already in operation.

1.3 Region of Study

Figure 1.1 below shows the region examined, along with all of the monitoring sites in Alberta considered under this work. Passive monitoring stations locations are shown as open circles, continuous monitoring stations as inverted open triangles, and the different Airsheds are indicated by the different colours of the station symbols. The Wood Buffalo Environmental Association (WBEA) and Lakeland Industrial Community Association (LICA) sites are shown in more detail later in this report (Figure 3.1 and Figure 3.6).

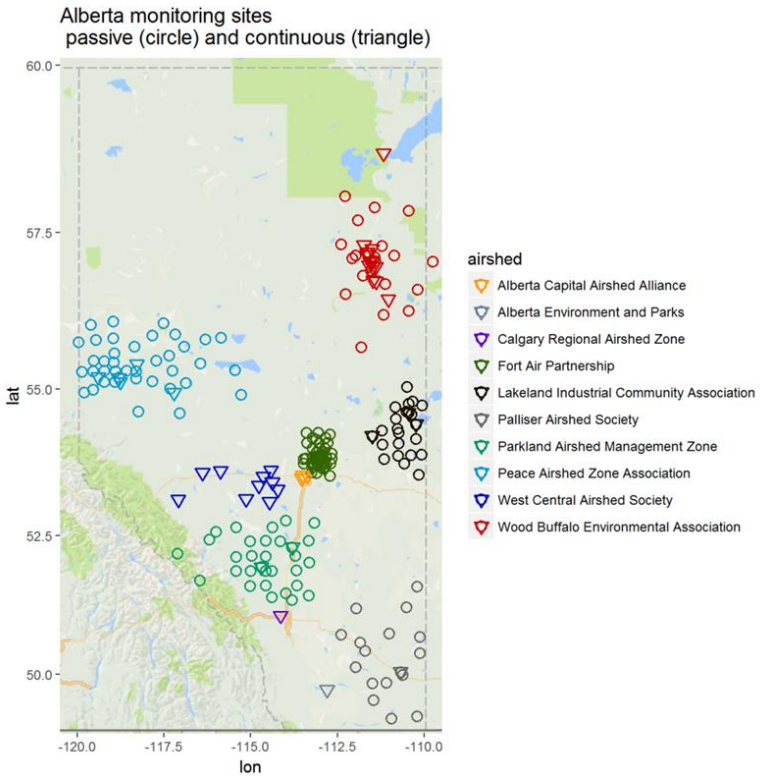


Figure 1.1 Air quality monitoring stations network in the province of Alberta

2 Methodology

2.1 Overview

There are three main components to the work reported here. The first of these was the collection, and quality control and assurance, of the available monitoring network data. The second and third stages relate to the analysis of that data; the use of Kolmogorov-Zurbenko (KZ) filtering of the data and subsequent hierarchical clustering to determine the level of similarity between stations. Note that the 2015 Workshop recommendations driving this work considered only one metric of station-to-station similarity (i.e., correlation, specifically the Pearson's correlation coefficient), though many other metrics may be used, and we have extended the analysis to also include the Euclidean norm.

We describe next the monitoring data and the procedures used for quality assurance and quality control, followed by an overview of the mathematical basis for the methodology (a more detailed description of the methodology appears in the Appendix).

2.2 Monitoring Data

2.2.1 Monitoring Network

Both continuous and passive samplers are used for assessing ambient air quality in the study region. Continuous sampling is carried out for regulatory compliance, and this requires high-temporal resolution in order to monitor short-term exceedances in highly variable concentrations of pollutants in ambient air. Passive sampling is carried out in order to determine monthly average ambient air concentrations of atmospheric compounds to determine long-term trends, assess potential exposure risks to ecological and understand spatial distribution of the measured pollutant.

The details of continuous monitoring methodologies for Ozone (O_3), Carbon Monoxide (CO), Nitric Oxide (NO), Nitrogen Dioxide (NO_2), Oxides of Nitrogen (NO_X), Particulate Matter with particle radius below 2.5 µm ($PM_{2.5}$), Sulphur Dioxide (SO_2), Non-Methane Hydrocarbons (NMHC), Total Hydrocarbon (THC), Total Reduced Sulfur (TRS) and Methane (CH_4) for Alberta monitoring networks are described elsewhere (AESRD, 2014); here we provide an overview via the minimum performance characteristics presented in Table 2.1. The majority of the Alberta passive monitors for NO_2 and SO_2 were developed by Maxxam Analytics Inc. (Tang et al., 1997; Tang et al., 1999; Tang, 2001), with exception of those employed by the Palliser Airshed Society (PAS), where the sampling program made use of a Multi-Gas Passive Sampler until May 2014, when it was replaced with the Radiello sampler tube (PAS, 2016). The underlying operating principle of these types of passive sampler is the collection of gas molecules by diffusion onto a collection medium coated with a chemical having specific affinity to the atmospheric compound of interest. The diffusion rate is controlled by pore size of the diffusion barrier, relative humidity, wind speed, and temperature. After collection, the exposed collection media are analyzed in the laboratory: SO_2 is analyzed via ion chromatography, and spectrophotometric and continuous flow analysis methods are used to estimate the cumulative NO_2. The time-weighted averaged concentration is calculated based on

the sampling period, the sampling rate and the collected cumulative mass for the sampling period. For a 30-day exposure period the detection limit for NO_2 and SO_2 samplers is 0.1 ppb. Material describing the validation of passive NO_2 and SO_2 samplers in air monitoring stations in Alberta may be found in Tang et al. (1997), Tang (1998), ARC (1998), Tang et al. (1999), and Brassard (2001).

Table 2.1 Minimum performance specifications and operating principles for continuous ambient air analyzers (AEP, 2016)

Criteria	O_3	NOx		$PM_{2.5}$ and PM_{10}
		Routine	Trace Level	
Operating Range (full scale)	0.5 or 1 ppm	0.5 or 1 ppm	0.2, 0.5 or 1 ppm	500 or 1000 µg/m3
Lower Detection Limit	1.0 ppb	0.5 ppb	0.05 ppb	4.8 µg/m3
Precision	1.0 ppb	0.5 ppb	0.05 ppb	2.0 µg/m3
Operating Principle(s)	Ultraviolet Photometry, Chemiluminescence	Chemiluminescence		USEPA Equivalent method
Criteria	**SO_2 Routine**	**THC/CH4/NMHC Trace Level**		**H_2S and TRS**
Operating Range (full scale)	0.5 or 1 ppm	0.1, 0.5 or 1 ppm		10, 20, 50 ppm
Lower Detection Limit	2.0 ppb	0.2 ppb		60 ppb
Precision	1.0 ppb[1]	0.2 ppb3		1% full scale
Operating Principle(s)	Ultraviolet pulsed fluorescence	Flame Ionization Detector (FID), Gas Chromatography/ FID, Oxidizer/FID		Ultraviolet pulsed fluorescence

[1] or 1% of reading, [2] or 0.5% of reading, [3] or 2% of reading

2.2.2 Monitoring Data Used for Analysis, Data Procedures

Continuous monitoring network data for the period from July 2013 through September 2014 for the species O_3, NO, NO_2, NOx, $PM_{2.5}$, SO_2, NMHC, THC, TRS and CH_4 were extracted from AEP archives, subjected to quality assurance and control procedures as defined below, and transferred to ECCC for analysis. The period was chosen in order to overlap with ECCC air quality model simulations covering the same time period, for cross-comparison under Stage 3 of the overall research project.

In order to examine the passive and continuous stations together, a further delivery of almost five years of monthly and bimonthly passive monitoring data for SO_2 and NO_2 were obtained from AEP records submitted by the operating Airsheds, as well as the corresponding five years of continuous monitoring data, for the period from February 2009 to December 2015. The Airsheds in Alberta are the organizations responsible for monitoring and reporting air quality to the public.

The first phase of the analysis of the one-year record of multi-species continuous monitoring data was carried out jointly in consultation between AEP and ECCC and focused on procedures to deal with gaps in the data. The analysis methodologies employed here require continuous data records (i.e. no gaps in the time series of observations used for analysis). The observing network data may have gaps (missing data), entries which indicate numbers below the detection limit of the observing samplers, and stations may have come on or off-line during the time interval selected. Following published recommendations for data analyzed with the methodology used here (Solazzo and Galmarini, 2015), stations were rejected from the analysis if their hourly data records for the period selected had more than 10% of the total data for the year missing or any data gaps which were more than 168 consecutive hours in length. This resulted in a number of stations being rejected from the analysis. The lack of delivery of useable data for analysis may in itself be a potential consideration for network optimization. Consultation with AEP was carried out to determine and tabulate the reasons for data rejection as these are summarized in Table 2.2 and Table 2.3 below.

As can be seen from Table 2.3, most of the stations were rejected on the basis of incomplete records for the period studied, but a few stations' records for specific chemicals were rejected since they did not observe the given chemical, were new stations, or were offline during part of the period selected. For the stations rejected on the basis of incomplete information, the analysis cannot be carried out; hence, no conclusions are possible for these stations aside from their low level of useable data during the period studied.

For those continuous monitoring stations which were *not* rejected due to missing data, shorter gaps in the data records still needed to be filled. The methodology to fill-in the gaps of the observational data follows Solazzo and Galmarini (2015): (1) For data gaps of 1 to 6 hours duration, the nearest flanking valid data on either side of the gap along with linear interpolation was used for gap-filling; (2) For data gaps of longer duration (but less than the 168 hour consecutive limit for data record rejection discussed above), the annual average of the non-gap data were used to fill the remaining gaps.

As a test of the second stage of this procedure on the clustering results, a variation was carried out wherein the longer gaps were filled using the average of the same amount of missing days both before and after the gap. No substantial difference was found between the resulting clusters in the subsequent analysis.

Table 2.2 Stations rejected from the continuous monitoring analysis (hourly values), grouped by chemical species, and the criteria which resulted in rejection (more than 168 hours missing, or more than 10% total data missing (90 percentile))

Parameter	NO_x	NO_x	O_3	O_3	SO_2	SO_2	$PM_{2.5}$	$PM_{2.5}$
Criteria	**168**	**90P**	**168**	**90P**	**168**	**90P**	**168**	**90P**
Station ID	1063	1056	1250	1056	1036	1056	1225	1168
Station ID	1029	1488	1071	1488	1071	1488	2000	1172
Station ID	1071	1495		1495		1495	1171	1224
Station ID		1476		1476		1479		1036
Station ID		1174		1174		1164		1476
Station ID						1476		1250
Station ID						1174		1488
Station ID						1068		1174
Parameter	**TRS**	**TRS**	**CH_4**	**CH_4**	**THC**	**THC**	**NMHC**	**NMHC**
Criteria	**168**	**90P**	**168**	**90P**	**168**	**90P**	**168**	**90P**
Station ID		1476	2001	1495	2001	1068	2001	1495
Station ID		1174	1049	1476	1049	1476	1049	
Station ID					1172	1052		
Station ID						1488		
Station ID						1174		
Station ID						1495		

Table 2.3 Continuous monitoring station notes, detailing the cause for station rejection from the analysis.

Station	Notes
1029	NO_x data missing for 268 hours
1036	SO_2 data missing for 323 hours, $PM_{2.5}$ data completeness 88%
1049	THC data missing 257 and 113 hours affecting CH_4 and NMHC as well
1052	THC data completeness 59%
1056	Hinton NO_x, SO_2 and O_3 completeness 67-69%
1063	NO_x data missing for 414 hours
1068	SO_2 data completeness 85%
1071	O_3, SO_2 and NO_2 data missing for 169 hours
1164	SO_2 data completeness 56%
1168	$PM_{2.5}$ data completeness 87%
1171	$PM_{2.5}$ data missing for 169 hours
1172	$PM_{2.5}$ data completeness 83%,
1174	THC data missing for 204 hours
1224	McIntyre building not an ambient station
1225	$PM_{2.5}$ data missing for 193 hours
1250	$PM_{2.5}$ data completeness 84%, ozone data missing for 416 hours
1476	Lancaster is a new station
1479	$PM_{2.5}$ data completeness 32%,
1488	Wapasu is a new station
1495	Calgary Southeast offline for much of the study period was being relocated
2000	$PM_{2.5}$ data missing 212 hours
2001	THC data missing for 505 hours (affects NMHC and CH_4)

For the comparison between passive and continuous SO_2 and NO_2 observations, similar quality assurance and control procedures were applied. The hourly continuous station data records were subject to the same station rejection criteria and gap-filling procedures as for the August 2013 through July 2014 continuous dataset described above. Passive monitors nominally record either one-month or two-month averages, depending on location. One-month data were averaged to bimonthly data in order to have a consistent time interval for the dataset (a requirement for the analysis). If one of the two-monthly values being averaged was missing from the original data, that bimonthly average was also treated as missing data. The resulting set of bimonthly data for all passive stations was then examined for completeness where stations which had greater than 25% missing data over the five year period were rejected from the subsequent analysis. This rejection criterion is less stringent than that applied to continuous data but it was set to achieve a balance between including monitoring sites with most complete data and attaining good spatial coverage. For example, an inclusion criterion of less than 10% of the data missing would have reduced the number of SO_2 passive sites included in the analysis from 52 sites to 18 sites and NO_2 passive sites from 39 sites to 18 sites. Table 2.4 and Table 2.5 summarize which continuous stations measuring NO_2 and SO_2, respectively, were rejected from the analysis and the cause for the rejection; Table 2.6 and Table 2.7 summarize which continuous stations measuring NO_2 and SO_2, respectively, were rejected from the analysis. The missing data were gap-filled using the averages for the given station for the remainder of the 5 year time period, in order to provide a contiguous time record for these stations (a requirement of the analysis). The gap-filled continuous data for the 5 year period were then averaged to the same bimonthly intervals. The averaging to a common bimonthly interval was done in order to allow the passive and continuous monitors to be analyzed together as a group, as will be described in more detail later in this report.

Table 2.4 Stations rejected from the NO2 continuous monitoring analysis (bimonthly averages), data completeness (more than 7 months missing, or more than 25% total data missing (75 percentile)), and detailing the cause for station rejection from the analysis.

Station	Data completeness (%)	Note
Albian Mine Site	1	Discontinued February 2009
Calgary Central-Inglewood	10	Data available from April 2015
Calgary East	33	Discontinued April 2011
Calgary Southeast	23	Data available from Nov 2015
Edson	55	Data available from Dec 2011
Firebag	14	Data available January 2015
Hightower	44	Discontinued July 2012
Hinton	25	Data available from Nov 2013
Lancaster	28	Data available from Nov 2012
Station 401	17	Discontinued March 2010
Stony Mountain	5	Data available from August 2015

Wagner	1	Discontinued January 2009
Wapasu	28	Data available from Dec 2013
Woodcroft	35	Data available from June 2013

Table 2.5 Stations rejected from the SO_2 continuous monitoring analysis (bimonthly averages), data completeness (more than 7 months missing, or more than 25% total data missing (75 percentile)), and detailing the cause for station rejection from the analysis.

Station	Data completeness (%)	Note
Albian Mine Site	1	Discontinued February 2009
Calgary East	33	Discontinued April 2011
Calgary Southeast	23	Data available from November 2015
Falher	23	Data available from May 2014
Firebag	14	Data available January 2015
Hightower Ridge	44	Discontinued July 2012
Hinton	25	Data available from Nov 2013
Lancaster	29	Data available from Nov 2012
Scotford 2	71	Discontinued April 2014
Stony Mountain	6	Data available from August 2015
Wagner	1	Discontinued January 2009
Wapasu	28	Data available from December 2013
Woodcroft	35	Data available from June 2013

Table 2.6 Stations rejected from the NO_2 passive monitoring analysis (bimonthly averages), and data completeness (25% total data missing (75 percentile) over the five year period).

Station	Data completeness (%)
192/22X	45
Airdrie	45
Arrowwood	45
Banff	45
Bay Tree	67
Bear Lake	69
Boone Creek	64

Bragg Creek	43
Calgary Applewood	40
Calgary East Village	40
Calgary Elbow Wetlands	43
Calgary Fish Creek	43
Calgary Metis Trail	43
Calgary Nose Hill	43
Calgary Pumphouse	43
Calgary Shepherd	40
Canmore	26
Clairmont Lake	67
Claresholm	43
Clouston Creek	64
Cochrane	26
Connemara	43
Crooked Creek	62
Crowfoot Crossing	43
Crowsnest Pass - Allison Creek Road	2
Crowsnest Pass - Bellevue	2
Crowsnest Pass - Blairmore Ranger Stn	2
Crowsnest Pass - Coleman North	2
Crowsnest Pass - Coleman Valley Floor	2
Crowsnest Pass - Crowsnest	2
Crowsnest Pass - Frank Slide	2
Crowsnest Pass - Lundbreck	2
Crowsnest Pass - Macload St.Kettle Creek	2
Crowsnest Pass - Pincher Creek Airport	2
Deer Mountain	60
Delacour	26
Eaglesham	60
East Prairie	26
FAP-01	48

FAP-02	48
FAP-03	48
FAP-04	48
FAP-05	43
FAP-06	43
FAP-07	48
FAP-08	48
FAP-09	45
FAP-10	40
FAP-21	29
FAP-33	45
FAP-34	48
FAP-35	48
FAP-36	48
FAP-38	48
FAP-39	48
FAP-40	45
FAP-41	48
FAP-47	45
FAP-51	48
FAP-53	48
FAP-58	36
FAP-59	10
FAP-60	26
FAP-62	26
Fitzsimmons	62
Flyingshot	67
Foster Creek	71
Fourth Creek	57
Frog Lake	76
Gift Lake	55
Gleichen	43

Gordondale	67
Grand Prairie I	67
Granum	24
Guy	64
High Prairie	67
Highwood Inn	40
Hythe	69
Jean Cote	62
Jumping Pound	26
Kananaskis Village	38
Karr Creek	57
Kinuso	52
Lake Louise	43
Langdon2	17
Langdon	24
Little Smoky	64
Lomond	36
Lyalta	43
McDougall Church	43
McLellan	62
Medley-Martineau	64
Mossleigh	43
Namaka	26
NW Border	24
Okotoks	43
Peacock	24
Peavine	52
Pinto Creek	67
Poplar	64
Portable Passive sample	19
Primrose	81
Puskwaskau	52

Rosebud	40
Saddle Hills	67
Sand River	45
Shaftesbury	52
Silver Valley	57
Spirit River	64
Stavely	36

Table 2.7 Passive monitoring station notes: causes for station rejection from the analysis Stations rejected from the NO_2 continuous monitoring analysis (bimonthly averages), and which criteria resulted in rejection (more than 1 months missing, or more than 25% total data missing (75 percentile)).

Station	Data completeness (%)
Bay Tree	69
Bear Lake	67
Boone Creek	67
Burnt Lake	71
Clairmont Lake	69
Clouston Creek	67
Crooked Creek	57
Deer Mountain	52
Eaglesham	64
FAP-01	71
FAP-04	71
FAP-07	74
FAP-08	74
FAP-10	71
FAP-11	71
FAP-12	64
FAP-18	74
FAP-23	71
FAP-24	67
FAP-26	74

FAP-27	74
FAP-29	74
FAP-30	69
FAP-31	67
FAP-32	69
FAP-33	69
FAP-37	74
FAP-38	71
FAP-39	71
FAP-42	74
FAP-43	71
FAP-45	74
FAP-48	74
FAP-49	74
FAP-51	71
FAP-52	74
FAP-53	74
FAP-54	69
FAP-57	67
Fishing Lake	69
Fitzsimmons	64
Flyingshot	67
Foster Creek	69
Fourth Creek	69
Gift Lake	52
Gordondale	71
Grand Prairie I	64
Guy	67
High Prairie	62
Hythe	69
Jean Cote	64
Karr Creek	43

Kinuso	57
Little Smoky	64
McLellan	62
Medley-Martineau	62
Muriel-Kehiwin	71
Peavine	71
Pinto Creek	52
Poplar	64
Puskwaskau	45
Saddle Hills	71
Shaftesbury	62
Silver Valley	69
Spirit River	64
Steeprock Creek	64
Sunset House	69
Sylvester	43
Valleyview	67
Wanham	69
Wapiti	62
Webber Creek	71
Wembley	55
Woking	69
Wolf Lake	55

2.3 Methodology for Station Data Analysis: Associativity Analysis

The *Workshop on Long-Term Air Monitoring Network Optimization* discussed in section 1.1 made reference to “Correlation analysis”, the methodology used here is more broadly known as “Associativity Analysis” or “Dissimilarity Analysis”, of the sub-type known as “Hierarchical Clustering” with metrics of evaluation being (a) 1-R where R is the Pearson correlation coefficient, and (b) the Euclidean distance. The methodology is based on prior work by Solazzo and Galmarini (2015) and others referenced therein.

For the analysis of multi-species continuous monitoring data, the (quality assured and controlled, gap-filled) hourly data were time-filtered to remove short-time-scale variations, using a moving average approach known as the KZ filter (Zurbenko, 1986). This resulted in four sets of observation-based time series for subsequent analysis; the original QA/QC gap-filled hourly datasets, and three additional datasets, which have had time variations less than a day, a week, and a month, removed. The subsequent analysis examined may thus examine the effect of each of these different time scales, to determine the extent to which patterns in the relationships between the stations are strengthened or weakened when shorter duration variation in the signal is removed.

In the second stage of the work, common for both the multi-species continuous and the bimonthly continuous + passive analyses, the associativity analysis known as hierarchical clustering was applied to the datasets for the stations, using two metrics for the degree of station associativity; correlation and the Euclidean distance (described in more detail below). The continuous stations with multispecies data were analyzed at the different time scales using this approach, for each of the four sets of filtered data. The combined five year record of passive and continuous bimonthly data were also analyzed using hierarchical clustering, without the *a priori* KZ filtering step (since the data themselves were already long-term averages).

The mathematical basis of both KZ filtering and hierarchical clustering are described in detail in Appendix 1. Here, we give a summary overview of the main points of the analysis. The KZ filter is a means for removing smaller time scales from a time series, based on an iterative moving average over a specific time window. The removal of high frequency variations of the data shows the relative influence of each of those time scales on the data. For example, data may have a large degree of variation every hour, but an underlying daily or weekly variation which may be of interest in analyzing the observations. The filtering allows these different time scales to be isolated and analyzed separately, hence gaining more information about the time variation of the data in a given analysis. The combination of the number of times the moving average is applied, and the duration of the averaging window, determines the time scales which are removed from the time series. Different combinations were used to filter out time scales: in this study, time scales smaller than daily, weekly, and monthly were removed from an initial time series of hourly data. The station data resulting from each level of filtering may then be cross-compared using hierarchical clustering, described below. We note here that we use the KZ filter in its original configuration, as a “low-pass” filter rather than as a “band-pass” filter. A “band-pass” filter is the difference between two low pass filters. The latter methodology was examined in Solazzo and Galmarini (2015); we found that the band-pass configuration performed poorly for distinguishing shorter time scales in numerical tests (described in more detail in Appendix 1, which also contains the mathematical details of KZ filtering).

It should be noted that *time-filtering* and *averaging* do not provide the same information. In the case of low-pass time-filtering, the higher frequency variation above some frequency is *removed* from the time series, while in the case of averaging, that information is added to the average. For example, if a plume with very high concentrations lasts three hours, then the daily average of the hours for the day containing that data will still be affected by that “spike”. Filtering of the data to remove the time scales of less than a day means that the effects of such spikes will be removed from the resulting time series. The average of the concentrations for a month with a few such events will include the effects of the events in the average,

while the time filtered data will have them removed. The methodology used here looks at those underlying time scales by filtering them out in successive stages, hence providing information on the time scales at which most of the variation occurs. In a correlation analysis, for example, station records which correlate with unfiltered data yet do not correlate when all time scales less than a month are removed, show that the variation resides within the time scales of less than a month.

Hierarchical clustering is a well-established method to determine the inherent or natural groupings of datasets, and/or to provide a summarization of data into groups. The grouping is done on basis of *similarities* or *dissimilarities*. Here, we will discuss the data in terms of their *dissimilarity*. The first step for hierarchal clustering is to choose a metric to describe how different (how dissimilar) a pair of data records (time series) are from each other, and calculate that metric for all possible pairs of the time series. After the level of dissimilarity has been calculated for each station with respect to every other station, the resulting dissimilarities are compared to each other and combined in the following procedure:

a) The two station records with the lowest level of dissimilarity are identified (i.e., the station records which are most "like" each other with respect to the metric of dissimilarity being used). This combination of stations becomes the first "cluster" of the analysis (i.e., clusters are groups of stations identified as being the most similar based on the dissimilarity metric).
b) The dissimilarities between this new cluster of two station records, and the remaining station records of the dataset, are then calculated. Here, the averages of the dissimilarity metric values between the two station records making up the new cluster, and each remaining station records, were calculated to describe these dissimilarities. This approach is known as the "general averaging method".
c) The dissimilarity values of the remaining station records, along with that of the new cluster, are examined again, and the most similar combination is identified. This combination may be between as of yet un-clustered station records, or between a station record and the new cluster. Once again, the general averaging method is used to combine the two.
d) The process of adding station records to existing clusters and/or combining clusters is repeated until all the station records have been clustered. The values of the metric used for dissimilarity as each new cluster is formed, along with the order in which the station records and clusters combine at each stage, are tracked.
e) Once all of the station records have joined a cluster, the tracked information (the order in which the station records combined with others and with clusters, and the level of the metric at which they combined) are used to generate diagrams called "dendrograms" which show the dissimilarity relationships between the stations. The relative ranking of the station records according to the dissimilarity level at which they join clusters show the relative dissimilarity (and hence similarity) between station records.

The analysis thus results in two main products: dendrograms showing the similarity relationships in detail, and tables of relative rankings of the (dis)similarity between station records. The data with the most similar records are potentially the most redundant, identical records being the extreme case.

The choice of a metric to describe the degree of dissimilarity between two stations is a crucial one, thus calculated similarities are only with respect to that specific metric. Different metrics may result in different rankings of stations on the degree of dissimilarity – hence the inferred level of potential redundancy also depends on the metric employed. Here, we have examined the dissimilarities resulting from two different metrics, and contrast their results in our analysis. The first of these metrics is "1-R", that is unity minus the Pearson correlation coefficient, the latter being the correlation coefficient between the two time series

being compared. This metric has been used in the past in dissimilarity analyses of air pollution observations (Solazzo and Galmarini, 2015; Yan and Wu, 2016). The second of these metrics is the Euclidean “distance”, the square root of the sum of the squares of the differences between the two time series at every value of the time series. The magnitude of the Euclidean distance, being a summation, will thus depend on the number of entries in the time series. Both of these metrics are used extensively within hierarchical clustering algorithms, and appear in texts on the methodology (e.g. Johnson and Wichern, 2007; Hastie *et al*., 2009; Næs *et al*., 2010). Appendix 2 contains a more detailed description of the hierarchical clustering methodology employed in this work, and the mathematical details of the metrics chosen for computing the dissimilarity between the time series.

Our reasoning in making use of the two dissimilarity metrics rather than 1-R alone is as follows, using a few examples. In many air pollution applications, one might expect pairs of stations to be aligned at different distances downwind from emission sources – the stations may thus be highly correlated, despite the increased dilution which might be expected with further distance from the source. High correlation alone may thus be an insufficient means by which to judge redundancy of station data, since the decrease in concentration with the distance from the sources will not feature into the analysis. However, the time series of two stations may also be very similar in magnitude but may be poorly or even anti-correlated due to being impacted by sources with emissions which vary differently over time. We therefore use both metrics in our analyses, noting that station time series pairs are the most similar, and hence potentially the most redundant, when both their 1-R and the Euclidean distance rankings for the stations are relatively low.

2.3.1 Dendrograms

As noted above, the results of hierarchical clustering may be displayed using specialized diagrams called “dendrograms”. Dendrograms show the pattern of linkages between the data series while clustering occurs, as well as their level of dissimilarity. Dendrograms thus resemble the root system of a tree, with the most similar stations forming the lowest level of the smallest roots, and the two least similar clusters being linked at the top of the diagram as the trunk of the tree. Vertical lines on the dendrogram represent the difference in the level of dissimilarity between consecutive stages of clustering; the horizontal lines show which time series or clusters of time series have been linked at a given level of the dissimilarity metric. A simple example of the construction of a dendrogram follows, in order to allow the reader to better interpret the subsequent results.

In this example, data for three different hypothetical stations are collected, and to measure their level of dissimilarity, the values of 1-R are calculated between each of the pairs of station records (Figure 2.1; the supporting tables give the values of the dissimilarity metric, 1-R). The data from stations 1 and 2 have a 1-R value of 0.5, stations 1 and 3 have a 1-R value of 0.4, and stations 2 and 3 have a 1-R value of 0.1. The lowest level of dissimilarity is thus between stations 2 and 3, and they are combined to become the first cluster, at a 1-R level of 0.1 (stations 2 and 3 are joined by a horizontal line in the dendrogram of Figure 2.1). The averages of 1-R between this cluster and the other stations are then calculated; in this case (0.5 + 0.4)/2 = 0.45, second table of Figure 2.1. Stations 2 and 3 thus cluster at 1 – R of 0.1, and the remaining station, 1, clusters at 1-R of 0.45. The second horizontal line of the diagram portion of Figure

2.1 shows the connection between the initial cluster between stations 2 and 3 and the final cluster with station 1. The result is a 3 station dendrogram (Figure 2.1).

Station 1-R	1	2	3
1	0	0.5	0.4
2	0.5	0	0.1
3	0.4	0.1	0

Station 1-R	1	2,3
1	0	0.45
2,3	0.45	0

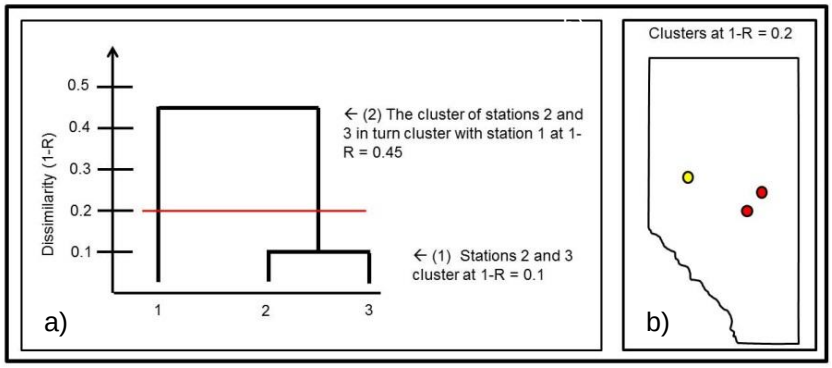


Figure 2.1. Example clustering of three stations: supporting tables to establish the dissimilarity between the stations (left); a) dendogram and b) map of spatial distribution of clusters.

Dendrograms for the Euclidean distance are similar in appearance to those for 1-R, the vertical axis of the dendrogram becoming the “distance” in concentration between the station records of the records being clustered.

If then we would assume a level of dissimilarity of 1-R = 0.2 (i.e. a correlation level 0.8) and draw a line over the dendogram at such level (red line, Figure 2.1), the clusters which fall below that line have a greater degree of similarity than the clusters or station records which join above that line. The locations of the stations that are included within clusters for a given level of dissimilarity may be then displayed on a map; the stations may be colour-coded according to their cluster of which they are a part. The resulting maps (e.g., Figure 2.1(b)) show the spatial relationships between station records which have a given level of dissimilarity. In the above example at correlation level 0.8 (1-R = 0.2) there are two sets of clusters: one comprised station 1, and the other comprised stations 2 and 3.

Stations may be *ranked* according to their *degree of dissimilarity* based on the level at which they join a cluster. In the above example, stations 2 and 3 join at the 1-R level of 0.1, and station 1 joins at the 1-R level of 0.45, and this may be displayed in Table 2.8 as follows:

Table 2.8 Relative ranking of stations 1, 2 and 3 based on the dissimilarity metric 1-R.

Station	1-R
1	0.45
2	0.10
3	0.10

The tables of relative dissimilarity place the most similar stations at the *bottom* of the table – Table 2.8, above, shows that the data records of stations 2 and 3 are the most similar. Also note that the two stations at the bottom of the table are linked to each other – the dendrogram must be used to determine which stations link with which clusters, in more complicated cases with more stations.

2.4 Choice of Stations to Cluster – Comparison of Networks versus Comparison Within Networks

The data analyzed here were collected by the Airsheds in Alberta (Figure 1.1). The data may be analyzed on a *provincial* basis, regardless of the source of information: as noted by Solazzo and Galmarini (2015), this may highlight useful information, such as clusters which may represent a lack of uniformity in measurement procedures across different jurisdictions. At the same time, conclusions of this nature must be drawn with care, since a common set of clusters within a given reporting jurisdiction may also represent sources that are unique to that region. Clusters of stations across geographically diverse locations may represent similar emitting processes occurring in those locations, while not necessarily indicating a physical link between the stations. These analyses are however useful, simply from the standpoint of identifying those similar processes occurring in the data. Redundancies, however, must be identified with these limitations in mind.

Clustering may also be carried out solely with records originating within a given Airshed (see Figure 3.2 and Figure 3.7 for examples) – these allow more specific estimation of potential redundancy, in the context of the expectation that the given stations are intended to measure primary and secondary pollutants originating from a physically nearby source or collection of sources, and hence the reasons for similarities across their data records may be less ambiguous.

2.5 Methodology Summary

Our analysis thus had the following steps:

- Following QA/QC procedures and for the data, KZ filtering was used on the continuous hourly datasets for the one-year period to remove variation corresponding to periods less than one day, one week, and one month.
- Time averaging of five years of hourly continuous data and monthly to bimonthly passive data for NO_2 and SO_2 were used to create bimonthly five year records for these two species.
- Hierarchical clustering was carried out:
    - For the passive and continuous bimonthly five year records, for the WBEA stations, the LICA stations, and the entire province of Alberta.
    - For the continuous data, for the entire province of Alberta
    - Using two different dissimilarity metrics, 1-R and the Euclidean norm, for each of these datasets.

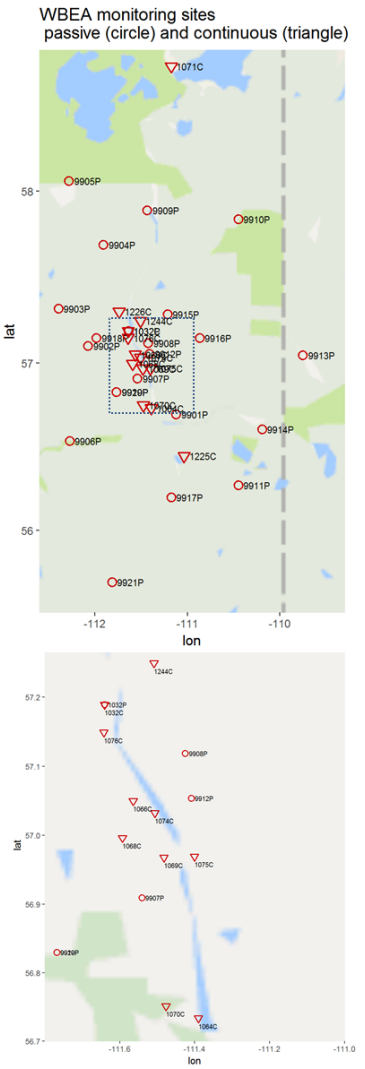
- Dendrograms were generated and clustering results were tabulated to show the relative ranking of similarity for each of these analyses and for each metric examined.

3 Applications of the Methodology

3.1 Associativity Analysis for WBEA: Five-year Combined Continuous and Passive Observations

The WBEA stations for which bimonthly NO_2 and SO_2 data were analyzed are shown in detail in Figure 3.1 for both the entire set of WBEA stations and for a zoomed-in view close to the main oil sands facilities in the Athabasca valley. Station Identification (ID) numbers in Figure 3.1 ending with a “P” or “C” letter suffix refer to passive and continuous monitors, respectively.

Stations were also grouped by AEP according to the dominant land-use around the station and/or the type of emissions sources nearby (Traffic, Point Source, Traffic and Point, Agriculture, Broadleaf Forests, Coniferous Forest, Grassland, Shrubland, Developed and Non-Specific) – these designations were initially used to colour-code station names in the resulting dendrogram analysis – however, no discernable pattern between the designations and the clustering could be observed, so that colour coding has not been retained here.



Station ID	Station Name
1032C	Fort McKay
1064C	Athabasca Valley
1066C	Mildred Lake
1068C	Buffalo Viewpoint
1069C	Mannix
1070C	Patricia McInnes
1071C	Fort Chipewyan
1074C	Lower Camp
1075C	Millennium Mine
1076C	Fort McKay South
1225C	Anzac
1226C	CNRL Horizon
1244C	Shell Muskeg River
1032P	Fort McKay-Bertha Ganter
9901P	AH3
9902P	AH8
9903P	BM10
9904P	BM11
9905P	BM7
9906P	JP101
9907P	JP102
9908P	JP104
9909P	JP107
9910P	JP205
9911P	JP210
9912P	JP212
9913P	JP213
9914P	NE10
9915P	NE11
9916P	NE7
9917P	SM8
9918P	WF4
9919P	AH7
9920P	R2
9921P	SM7

Figure 3. 1 Wood Buffalo Environmental Association's (WBEA) monitoring stations, located in the Athabasca oil sands region (bottom map: zoom over the blue box on upper map). Continuous stations are shown as inverted triangle, passive stations as circles.

3.1.1 NO2 Dissimilarity Analysis, WBEA Stations

The NO_2 dendrograms resulting from the use of 1-R as the dissimilarity metric is depicted in Figure 3.2, for the WBEA Athabasca oil sands region. Figure 3.2 (a) shows that the 1-R clustering follows two broad groups in the first branching of the dendrogram, taking place at correlation level of 0.41 (1-R = 0.59): on the right, a cluster composed of a set of passive monitoring stations (JP213/9913P, BM10/9903P, BM11/9904P, BM7/9905P, NE7/9916P, R2/9920P, JP212/9912P, JP205/9910P, JP107/9909P, NE11/9915P, NE10/9914P, SM7/9921P) and on the left, the remaining passive monitoring stations (AH8/9902P, JP102/9907P, JP104/9908P, Fort Mckay-Bertha Ganter/1032P, JP101/9906P, AH3/9901P, JP210/9911P, and AH7/9919P) clustered together with the continuous monitors. For this left branch, the continuous monitor at Fort Chipewyan/1071C separates out, followed by the remaining continuous

monitors cluster separating from the passive monitors at a correlation level of 0.68 (1-R=0.32). Fort Chipewyan/1071C is located far to the north of the other continuous monitors (see Figure 3.1, circled station), and far from the sources of emissions around which the other monitors are located. Thus, the NO_2 concentration record at Fort Chipewyan/1071C might be expected to be different the remaining continuous monitoring sites, all of which are in closer physical proximity to each other, and this explains the separation of Fort Chipewyan/1071C from the rest of the continuous stations early in the clustering. However, the separation of all of the continuous monitors from the passive monitors within the first three branches of the dendrogram reflects a systematic difference between the measurement methodologies employed in each case – that is, the performance of the two types of monitors is sufficiently different that they form different clusters. It should also be noted that while some of the passive monitors in this left branch (the ones most closely linked to the continuous monitors) are in close physical proximity to the continuous monitors (specifically, JP104/9908P, JP102/9907P, AH3/9901P, AH8/9902P, AH7/9919P), two others are more distant (JP210/9911P and JP101/9906P). The larger distance implies similar local source types, or high degree of uncertainty in the observations. Passive stations JP212/9912P and WF4/9918P meanwhile are relatively close to the group of continuous monitors, but do not cluster closely with them. These two passive stations are on the opposite side of the river valley from the continuous sites (hence local topographical features may modify the meteorological flow, isolating the stations from each other, and thus may play a role in the lack of clustering with nearby sites at this correlation level), but JP212/9912P clusters most closely with the group including JP205/9910P, JP107/9909P, NE11/9915P, NE10/9914P, which are considerably more distant from the sources. These groupings do not seem to follow proximity to the sources, and may indicate other processes aside from emissions proximity dominating the local concentrations at these sites, and/or low precision in the observations. The clusters at a given correlation level may be mapped to show their spatial distribution and gain further insight in the analysis. Figure 3.2(b) shows the clusters at a correlation coefficient of 0.75 (1-R=0.25), each cluster present at this level having been assigned a different colour. The clusters at this level are: Fort Chipewyan/1071C (a single member cluster), the remaining continuous monitors, and six clusters of passive monitors: a first group comprised of AH8/9902P, JP102/9907P, JP104/9908P, Fort Mckay-Bertha Ganter/1032P, JP101/9906P, AH3/9901P, JP210/9911P, and AH7/9919P, JP213/9913P as a single member cluster, a cluster between BM10/9903P, BM11/9904P, BM7/9905P, a cluster between NE7/9916P, R2/9920P; JP212/9912P as a single-member cluster; and a final group of passive monitors (JP205/9910P, JP107/9909P, NE11/9915P, NE10/9914P, SM7/9921P). This division between the sites at R=0.75 might indicate that the stations are monitoring very different sources, or air masses, since they cluster at a relatively high correlation level.

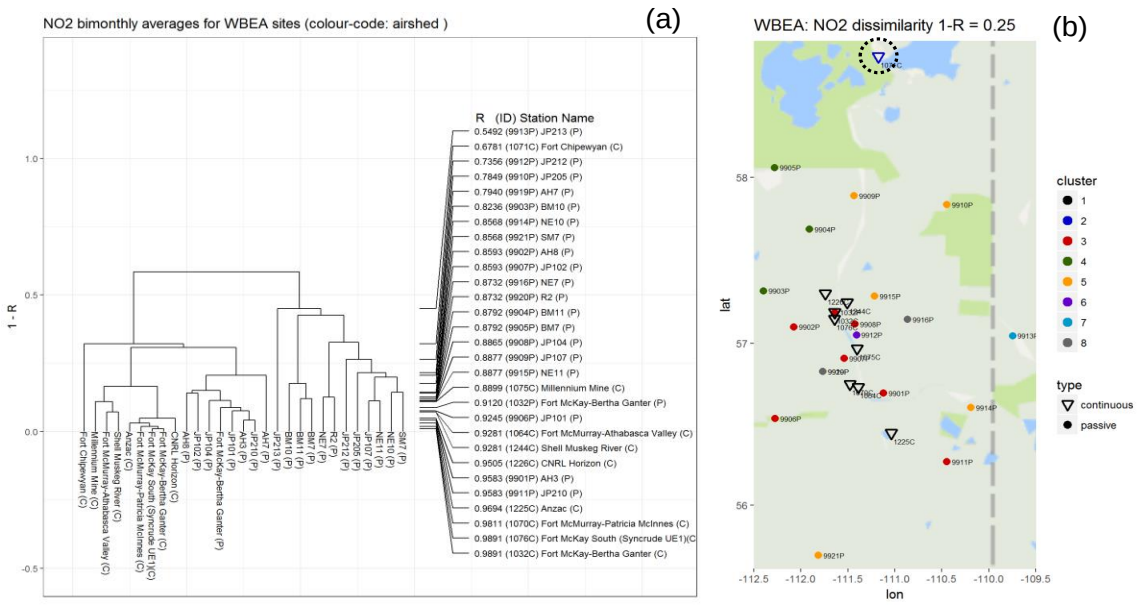


Figure 3. 2 (a) Dendrogram for passive and continuous bimonthly NO_2 averages considering 1-R as the dissimilarity matrix, for Wood Buffalo Environmental Association (WBEA). (b) Clustering of stations for 1-R=0.25

The dendrogram considering Euclidean distance as a metric to compute the dissimilarities (Figure 3.3 (a)) shows a very different pattern from the correlation analysis. For example, three of the continuous stations (Millennium Mine/1075C, Fort McMurray-Athabasca Valley/1064C, and Shell Muskeg River/1244C) are located to the north, directly within, and to the south of all of main NO_2 emitting region - and hence might be expected to have the greatest difference in their Euclidean distances. Figure 3.3 (b) shows the clusters which result at a Euclidean distance 25% of the maximum distance shown in the dendrogram (i.e. a Euclidean distance of 6 ppb): sixteen stations form a single cluster at this level, though their wide distribution of spatial locations both close to and far from the main emission region suggests imprecision in the observations, a similar low “background” level being observed at all stations, and/or very local conditions may play a role in the clustering for the passive stations. For example, passive station JP102/9907P clusters with continuous station Anzac/1225C at a Euclidean distance of 5.1 ppb, but not with continuous stations Patricia McInnes/1070C, or Athabasca Valley/1064C, which are located between JP102/9907P and Anzac/1225C. There may also be similar issues with the data from the continuous stations. For example, CNRL Horizon/1226C, Fort McKay-Bertha Ganter/1032C, Fort McKay-South/1076C and Fort McMurray-Patricia McInnes/1070C all cluster for Euclidean distances less than 7 ppbv – the first three stations are in relatively close proximity, and hence might be expected to cluster together, but the last is located on the far (south) side of the main emissions region from the first three. Also, despite being the close proximity between Fort McMurray-Patricia McInnes/1070C and Fort McMurray – Athabasca Valley/1064C, these stations do not cluster closely using the Euclidean distance metric, suggesting other factors at play in setting concentrations at these sites.

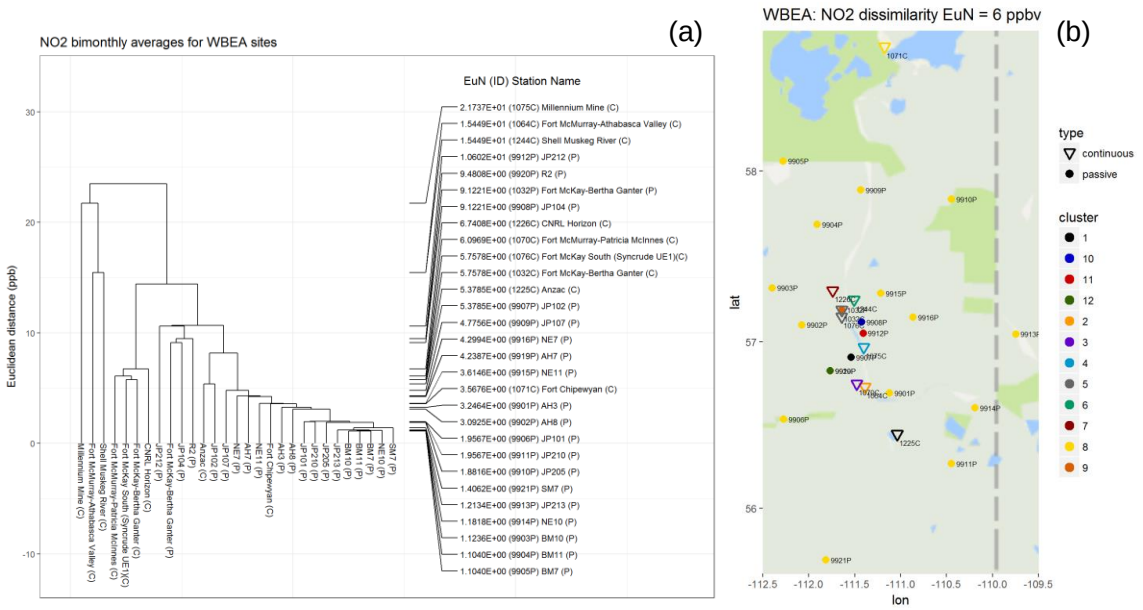


Figure 3. 3 (a) Dendrogram for passive and continuous bimonthly NO_2 averages considering Euclidean distance as the dissimilarity matrix, for Wood Buffalo Environmental Association (WBEA). (b) Stations colour coded for clusters with a Euclidean norm of 6 ppbv.

3.1.2 SO_2 Dissimilarity Analysis, WBEA Stations

The clustering results using the metric 1-R for SO_2 are not as clear as for NO_2; Figure 3.4(a) shows that the station data form clusters at higher dissimilarity values (lower correlation coefficients) for SO_2 than for NO_2 (compare to Figure 3.2(a)). Figure 3.4 (b),(c) show the spatial distribution of stations at 1-R values of 0.8, 0.5, and 0.3 (R=0.2, 0.5 and 0.7, respectively). The higher dissimilarity values in Figure 3.4(a) reflect the nature of the SO2 sources; unlike NO_2 (a large component of the emissions of which come from surface-based area sources such as the “heavy hauler” trucks at the mine sites), the SO_2 emissions are almost entirely from stacks (aka “point sources”). The low correlations may thus reflect the more time dependent nature of SO_2 emissions, as well as the dependence of the resulting downwind concentrations on meteorological variables throughout the atmospheric column, such as the atmospheric stability at the point of emissions (controlling plume rise), and the wind at different levels (controlling the downwind dispersion direction of the plumes). It is also worth noting that for stations rarely impacted by high concentration plumes, the background SO_2 may close to or below the detection limit of the sampler, reducing correlations.

As was seen NO_2, the first branching clearly separates two set of stations, this time, though, there is less of a clear distinction between passive and continuous monitors in the clustering, though many of the continuous monitors are part of a single cluster for 1-R < 0.57, and a subgroup of continuous and passive stations (Fort McKay-Bertha Ganter/1032P,1032C; Fort McKay-South/1076C; JP102/9907P; Mannix/1069C; and CNRL Horizon/1226C) are part of a single cluster for 1-R< 0.54. There is some degree of consistency with location; for example, Mannix and JP102 or continuous stations in Fort Mackay (1032C, 1076C, 1032C) are relatively close to each other spatially, and cluster with a higher

correlation coefficient (R =0.7 and 0.837), Shell Muskeg River/1244C is at times directly downwind of NE11/9915P (R = 0.68), as are Mildred Lake/1066C and passives JP104/9908P and JP212/9912P.

The SO_2 dendrogram for Euclidean distance is shown in Figure 3.5 (a), and the clusters for Euclidean distances of 5.0, 4.0, and 3.0 ppbv are shown Figure 3.5 (b), (c),(d), respectively. The dendrogram for SO_2 Euclidean distance (Figure 3.5 (a)) shows clustering taking place at smaller Euclidean distances compared to its NO_2 counterpart (see Figure 3.3), indicating a lesser degree of dissimilarity between SO_2 stations than NO_2 stations. The SO_2 nodes also have a smaller dynamic range from maximum to minimum Euclidean distance. However, some stations tend to cluster more closely than others – for example, Fort Chipewyan/1071C, NE10/9914P, BM7/9905P, and SM8/9917P all cluster together for a Euclidean distance of 3.33 ppbv (cluster 3 in Figure 3.5 (c)) , clustering further with stations such as Anzac /1225C, JP210/9911P, JP213/9913P at 4.7 ppbv (cluster 1 in Figure 3.5 (b)). All these stations have the common feature of being located a significant distance from the main emissions sites varying from 75 to 188 km from the Syncrude main stack. Figure 3.5 (b), (c) shows how the stations cluster regarding the distance to the main sources in the area. The analysis thus suggests that for SO_2, the Euclidean distances become more similar with increasing distance from the sources. This “makes sense” on an intuitive level, in that close to the emission sources, the plumes from the large stacks will be very distinct and very dependent on the local meteorological conditions – while further downwind, the plumes will be more dispersed, and have a greater chance of being sampled at more than one downwind site at the same time, with similar concentrations.

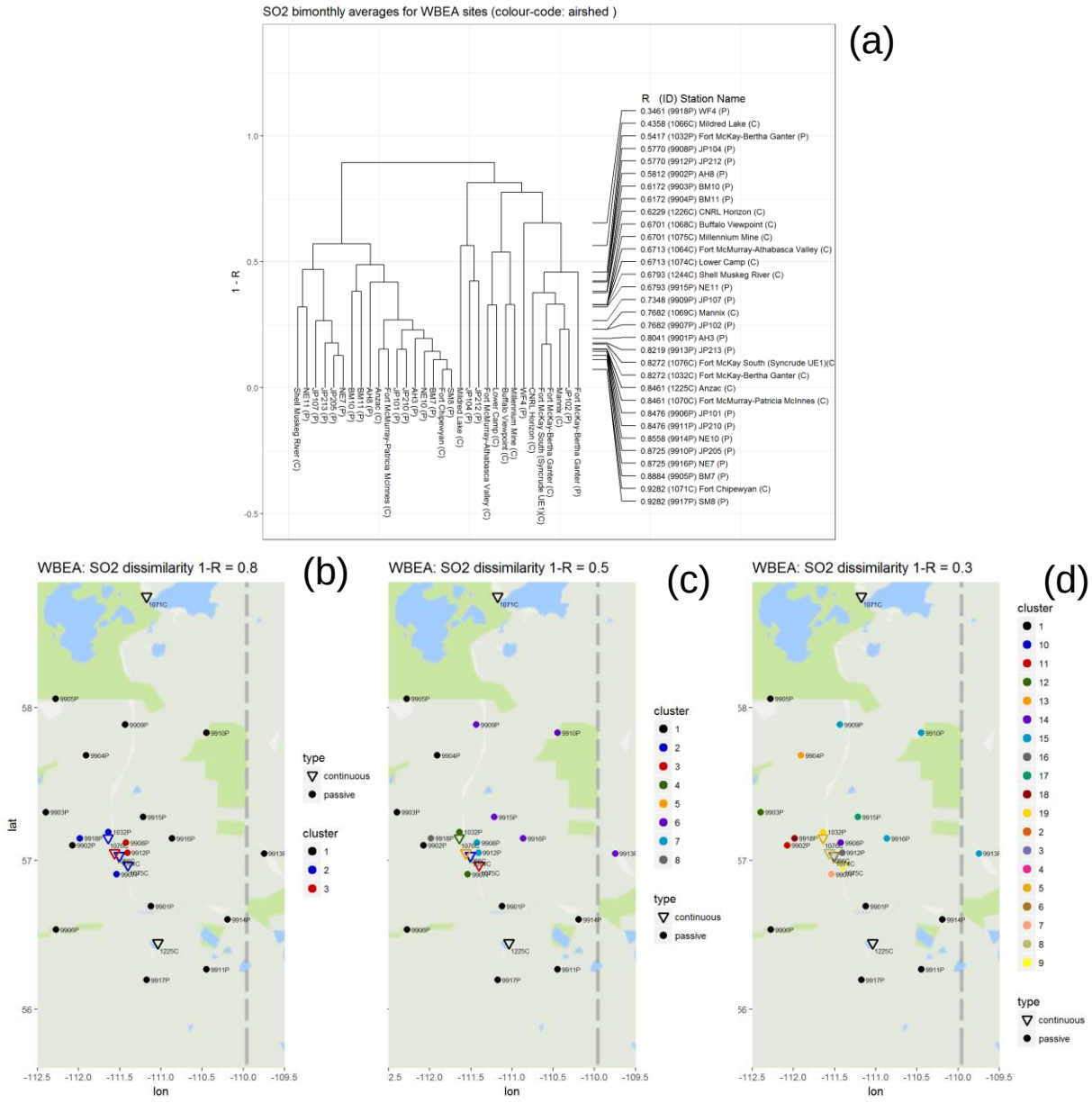


Figure 3. 4 (a) Dendrogram for passive and continuous bimonthly SO_2 averages using 1-R as as the metric to compute the dissimilarity matrix, for Wood Buffalo Environmental Association (WBEA). (b),(c),(d) Station clusters for 1-R of 0.8, 0.5 and 0.3, respectively.

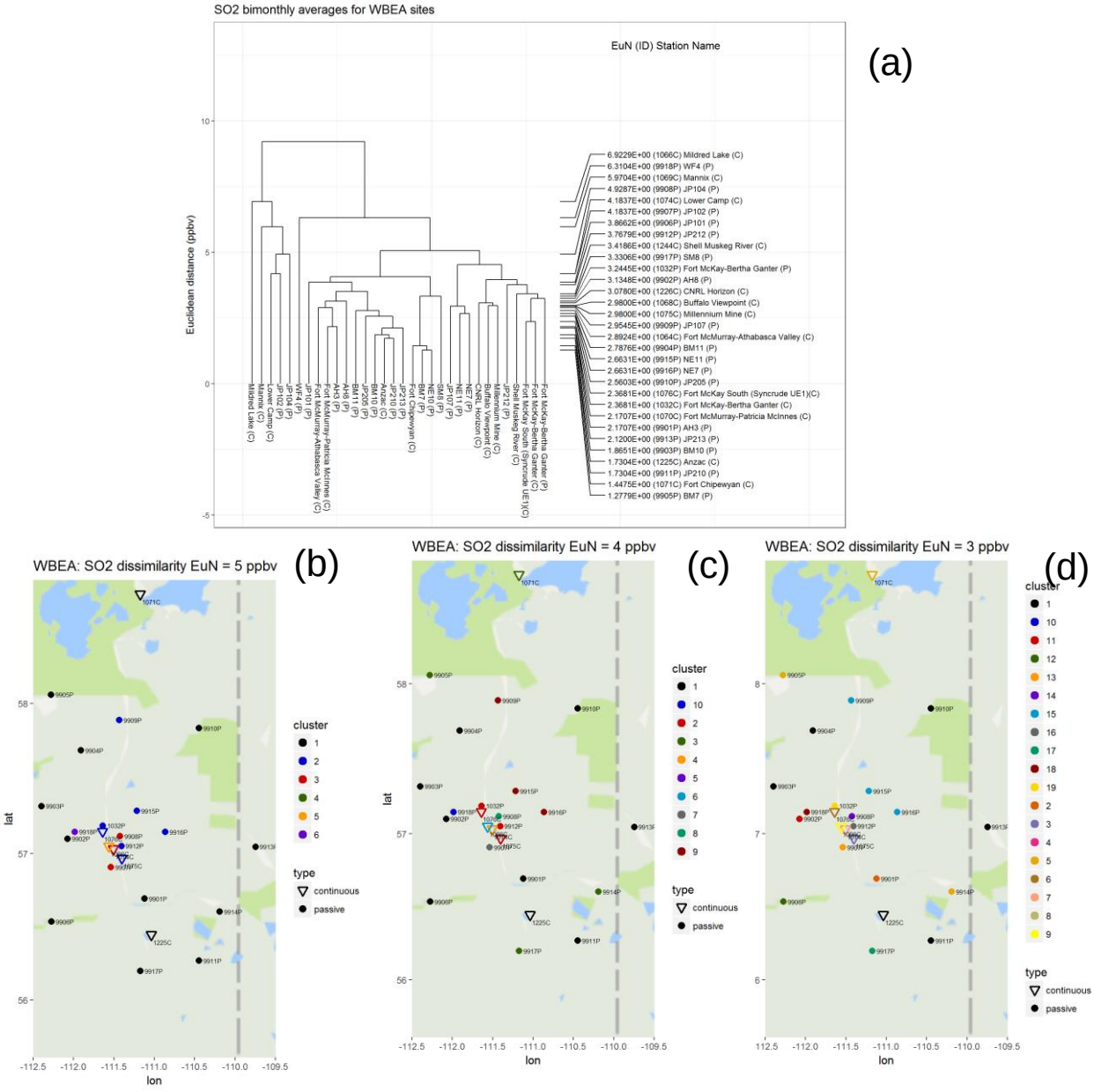
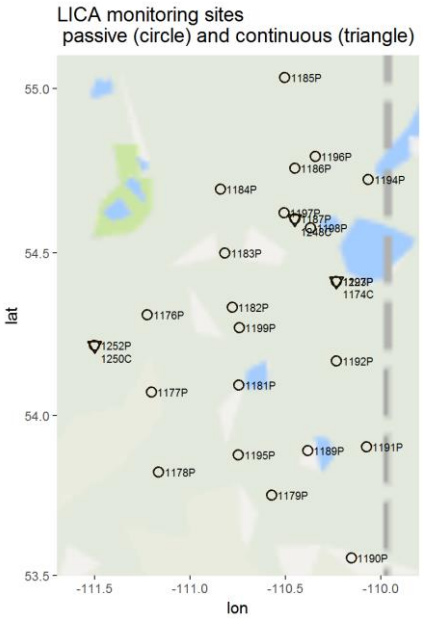


Figure 3. 5 (a) Dendrogram for passive and continuous bimonthly SO_2 averages using Euclidean distance as the metric to compute the dissimilarities, for Wood Buffalo Environmental Association (WBEA). (b), (c), (d) Stations colour-coded for Euclidean distances of 5.0 ppbv, 4.0 ppbv and 3.0 ppbv, respectively.

3.2 Associativity Analysis for LICA: Five-year Combined Continuous and Passive Observations

The station names and identification numbers for continuous and passive data operated by LICA are shown in Figure 3.6, below. One noticeable difference from the WBEA stations is the predominance of passive monitoring stations, with only three sites (Cold Lake South/1174C, Maskwa/1248C and St. Lina/1250C) having continuous monitors.



Station ID	Station Name
1174C	Cold Lake South
1248C	Maskwa
1250C	ST.LINA
1176P	Therien
1177P	Flat Lake
1178P	Lake Eliza
1179P	Telegraph Creek
1181P	Muriel-Kehiwin
1182P	Dupre
1183P	La Corey
1184P	Wolf Lake
1185P	Foster Creek
1186P	Primrose
1187P	Maskwa
1189P	Frog Lake
1190P	Clear Range
1191P	Fishing Lake
1192P	Beaverdam
1193P	Cold Lake South Passive
1194P	Medley
1195P	Fort George
1196P	Burnt Lake
1197P	Mahihkan
1198P	Hilda Lake
1199P	Town of Bonnyville
1227P	Cold Lake South Passive 2
1252P	St. Lina

Figure 3.6 Stations in Lakeland Industrial Community Association's oil sands region, located near Cold Lake, Alberta.

3.2.1 NO_2 Dissimilarity Analysis, LICA stations

The LICA NO_2 dendrogram using 1-R as a dissimilarity metric Figure 3.7(a) shows the first and second branching occur at correlation level 0.03 and 0.56 (1-R = 0.97 and 0.44), corresponding to passive stations Primrose/1186P and St. Lina/1252P, respectively. Additionally, the analysis shows that the passive and continuous monitors have sufficiently dissimilar time series that collocated monitors of each respective type do not fall within the same cluster. For example, the passive and continuous pair located at St Lina (1250C and 1252P) are collocated on the map scale used in Figure 3.7 (a), but do not cluster closely despite this co-location (we note also that previous work (Bari et al, 2014) suggested that passive NO_2 monitors tend to be biased low relative to collocated continuous monitors). These examples may also indicate the level of error in the observations, and/or specific events recorded at one station and not another. With regards to potential sources of error, we note that the time series for Primrose/1186P showed that it included a single isolated high concentration data point which does not appear in the data for the surrounding stations. However, St. Lina was chosen as an upwind site from the sources in the

LICA Airshed – so its tendency not to cluster with other stations may be related to its location. The methodology has identified the time series for these stations as being dissimilar from the others (and hence helps suggest potential QA/QC methodologies which could be used in the future). The underlying causes for that dissimilarity must be based on examining the time series and using local knowledge of sources and conditions.

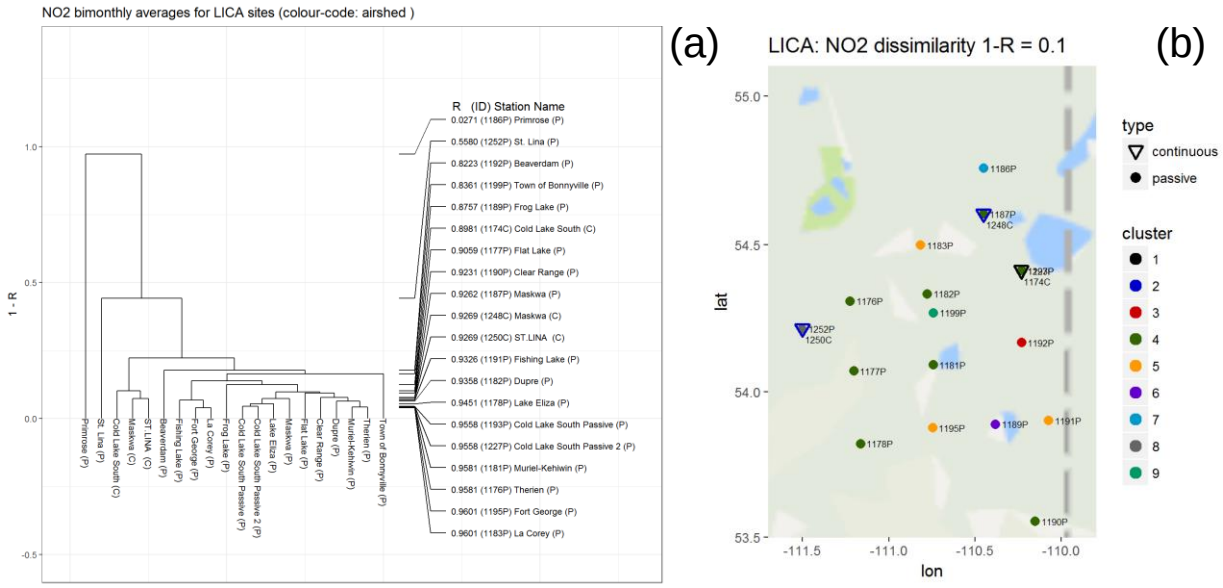


Figure 3.7 (a) Dendrogram for passive and continuous bimonthly NO_2 averages using 1-R as the dissimilarity matrix, for Lakeland Industrial Community Association (LICA). (b) Stations colour-coded by cluster for 1-R =0.1 (R=0.9). Stations with continuous monitors end in a "(C)", and stations with passive monitors in a "(P)".

Returning to Figure 3.7 (a), at the third branching, at correlation level 0.79 (1-R=0.21), the stations split, with the continuous monitors clustering on the left and the remaining passive monitors on the right. Within the cluster of three continuous monitors, correlations are very high; all are greater than 0.9 (1-R<0.10) and monitors Maskwa/1248C and St Lina/1250C correlate at above the 0.9. It is important to note that all three continuous monitors correlate highly despite their separation in space and in comparison to the proximity between the St. Lina passive and continuous monitors. The analysis thus suggests that continuous and passive monitors within this region are differing in the degree of similarity – the former being highly similar with respect to each other, the later having a lower degree of similarity with both the continuous monitors and other passive monitors. The larger "passive" cluster to the right of Figure 3.7 (a) comprises stations such as Beaverdam/1192P, Frog Lake/1189P, and Town of Bonneyville/1199P. These have very different land-use types according to records provided by AEP ("broadleaf forest", "shrub-land" and "developed", respectively), suggesting that for these locations, the land-use designations have little impact on correlation. There are two clusters of passive stations with a correlation level of over 0.9 (1-R=0.1): one comprised of Fishing Lake/1191P, Fort George/1195P and La Corey/1183P and the other including the remaining passive stations (Cold Lake South Passive/1193P, Cold Lake South Passive2/1227P, Lake Eliza/1178P, Maskwa/1187P, Flat Lake/1177P, Clear Range/1190P,

Dupre/1182P, Muriel-Kehiwhin/1181P, and Therien/1176P). Figure 3.7(b) shows the stations mapped with common colours within clusters for R=0.9, showing the locations of these highly correlated groups.

The NO_2 dendrogram using Euclidean distance as the dissimilarity metric (Figure 3.8(a)), shows passives Primrose/1186P and Town of Bonnyville/1199P forming the first two branches of the dendrogram at 15.6 and 10.4 ppb, respectively, followed by the continuous monitor at Cold Lake South at level 10.1 ppb. Unlike the 1-R metric, the Euclidean distance does not have the continuous monitors clustering closely, as was noted for the WBEA stations. The continuous stations for NO_2 in both WBEA and LICA cluster closely for 1-R, but poorly for Euclidean distance. This suggests that NO_2 observed at these stations co-vary but differ in magnitude, the former possibly indicating a similar time dependence for the NO2 sources and/or meteorology, and the latter indicating influence by similar sources, but differences in downwind magnitudes due to dispersion, transformation or deposition of NO_2. The remaining monitors start branching out in the dendrogram at a level of 6.5 ppb. A few passive monitors form a single cluster at levels lower than 2 ppb: Flat Lake/1177P, Fishing Lake/1191P, Lake Eliza/1178P and Muriel-Kehiwin/1181P. Figure 3.8(b), (c) show the stations colour-coded by clusters at Euclidean distances of 5 ppbv and 2.5 ppbv, respectively. As discussed in more detail later in this report, clustering of stations may sometimes result from being located where there is little influence from sources – this may be the case here; all four passives appear to be located in a similar (remote) environment.

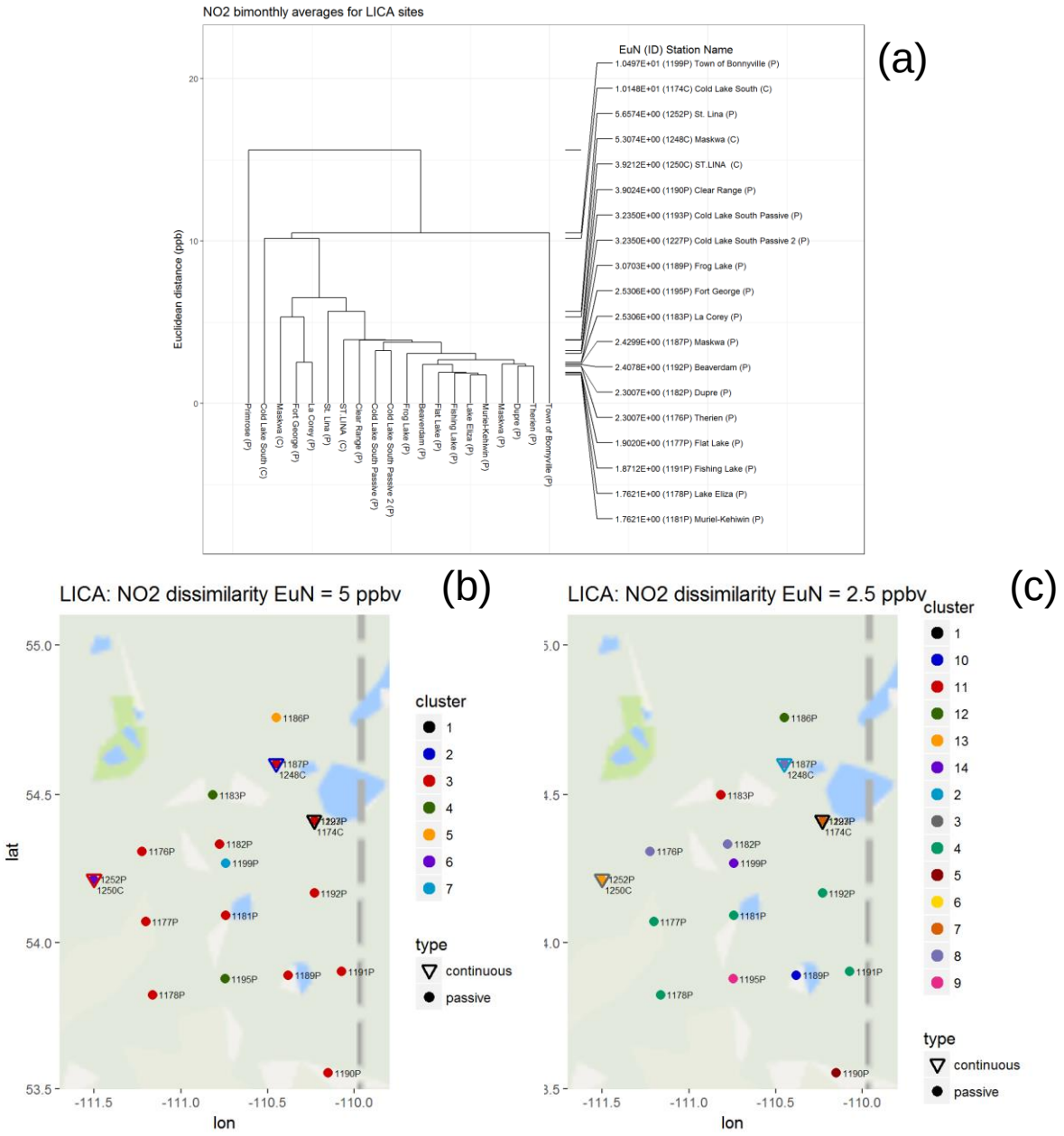


Figure 3.8 (a) Dendogram for passive and continuous bimonthly NO_2 averages using Euclidean distance as the dissimilarity matrix, for Lakeland Industrial Community Association (LICA). (b),(c) Stations colour-coded according to Euclidean distances of 5 and 2.5 ppbv, respectively. Stations with continuous monitors end in a "(C)", and stations with passive monitors in a "(P)".

3.2.2 SO_2 Dissimilarity Analysis, Cold Lake region

The SO_2 dendrograms resulting from the use of 1-R as the dissimilarity metric (Figure 3.9(a)) show a different pattern from those for NO_2, as all three continuous stations do not cluster together at as high levels of correlation as for NO_2. Town of Bonnyville/1199P branches out as a single cluster at correlation level 0.02 (1-R = 0.98) followed by the continuous station Cold Lake/1174P at correlation level 0.32 (1-R

= 0.68). The third bifurcation forms a cluster between Maskwa/1248C and St. Lina/1250C at a correlation level 0.41 (compared to R=0.90 for NO_2). The fourth bifurcation, at correlation level 0.51 (1-R = 0.49), divides passives Hilda Lake/1198P and Telegraph Creek/1179P from the remaining passives, followed by Maskwa/1187P and Primrose/1186P clustering at correlation level 0.79 (1-R = 0.21). Figure 3.9(b), presenting the stations colour-coded by cluster for R values of 0.7, shows Maskwa and Primrose cluster as cluster 6 and the remaining passives (aside from Town of Bonnyville/1199P) being represented as cluster 4. Figure 3.9 (c) shows some tendency of stations to cluster according to distance from the sources to the north-east of Cold Lake, with many stations further from these sources falling within a common cluster (cluster 4; Therien/1176P, Flat Lake/1177, Lake Eliza/1178P, Clear Range/1190P, Beaverdam/1192P and Fort George/1195P)). There are passives samplers clustering in two different clusters at correlation level 0.90 (from left to right): Fort George/1195P and Flat Lake/1177P; and Cold Lake South Passives 1 and 2 (1193P and 1227P, respectively). This analysis also shows that collocated passive and continuous monitors poorly correlate (Maskwa (1187P/1248C), Cold Lake South (1193P/1227P/1174C) and St. Lina (1252P/1250C)).

The SO_2 dendrograms which use the Euclidean distance as the dissimilarity metric (Figure 3.10(a)) show smaller magnitude differences between the stations, compared to NO_2 (compare vertical axes of Figure 3.8(a) and Figure 3.10(a)). The detection limit of SO_2 in the continuous monitors used here is 0.5 ppbv, while the passive detection limit is 1 ppbv – many of the stations are thus reporting values close to the detection limit, increasing their similarity for the Euclidean distance metric. Almost all the stations cluster at Euclidean distances of less than 2.5 ppbv distance, with exception of the continuous monitor St Lina/1250C and Cold Lake South/1147C, forming a single cluster at 3.9 ppbv and 3.4 ppbv, respectively. Passives Hilda Lake/1198P and Maskwa/1187P cluster together at 2.1 ppbv but are the first clusters branching out of the dendogram at a distance level of 5.3 ppbv. The collocated passive and continuous pair at St Lina (1252P/1250C), Maskwa (1248C/1187P) and Cold Lake South (1174C/1193P, 1227P) do not cluster with each other, again suggesting that the passive and continuous observations are not equivalent. Figure 3.10(b) shows the clusters resulting for a Euclidean distance of 1.0 ppb. Two different clusters of passive stations have a Euclidean distance within 1 ppb, one comprising of Cold Lake South Passives 1 and 2/1193P/1227P, La Corey/1183P, Beaverdam/1192P, Dupre/1182P, Fort George/1195P and Therien/1176P (cluster 4), and the other of Flat Lake/1177P, Clear Range/1190P and Lake Eliza/1178P (cluster 5).

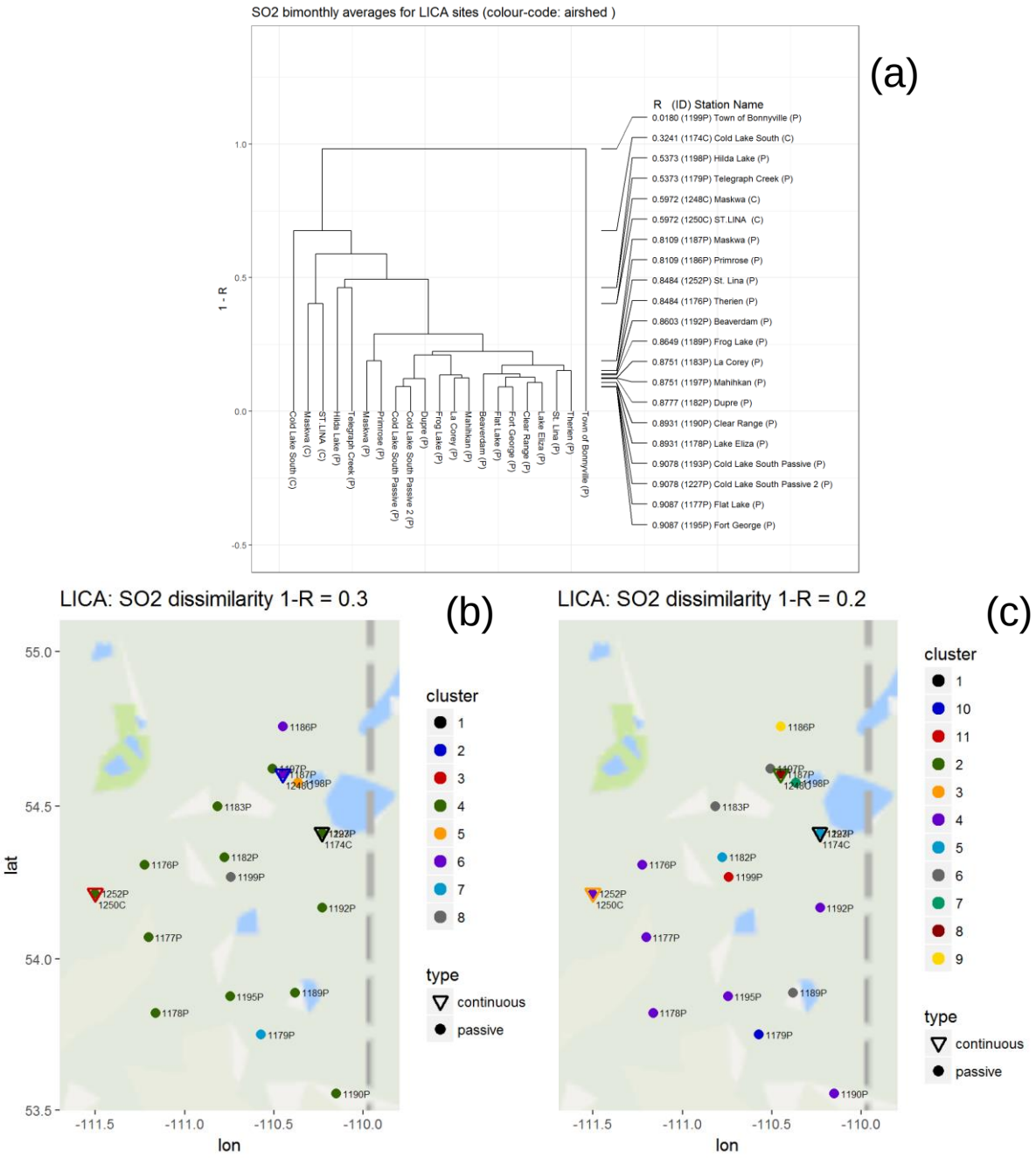


Figure 3.9 (a) Dendogram for passive and continuous bimonthly SO_2 averages using 1-R as the dissimilarity metric, for Lakeland Industrial Community Association (LICA). (b) Station locations colour-coded by R value for R=0.7 and R=0.8, respectively. Stations with continuous monitors end in a "(C)", and stations with passive monitors in a "(P)".

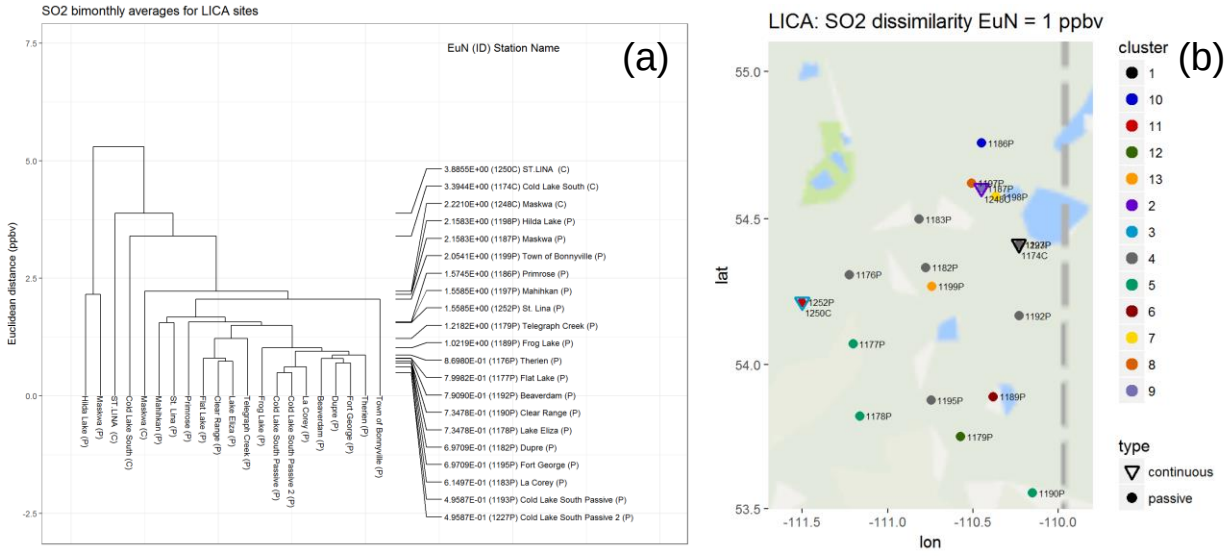


Figure 3.10 (a) Dendogram for passive and continuous bimonthly SO_2 averages using Euclidean distance as the dissimilarity metric, for Lakeland Industrial Community Association (LICA). (b) Stations colour-coded by cluster for a Euclidean distance of 1 ppb. Stations with continuous monitors end in a "(C)", and stations with passive monitors in a "(P)".

3.3 Associativity Analysis for Alberta: Passive and Continuous Bimonthly Observations

The same procedures for data selection and hierarchical clustering described above were applied to the five year record of bimonthly NO_2 and SO_2 observations for the entire province of Alberta. As before, station ID's contain a "P" or "C" letter, referring to passive and continuous monitors, respectively.

The NO_2 1-R metric dendrogram for the five year dataset, with station names colour-coded by the Airsheds, is shown in Figure 3.11. A prominent feature of this dendrogram is that clustering largely by Airshed can be seen for stations within the four Airsheds with the largest numbers of stations (PAS, Parkland Airshed Monitoring Zone (PAMZ), LICA, and WBEA (the latter is broken into subgroups, though large numbers of stations nevertheless cluster together). This clustering might be expected for stations with internally similar measurement procedures and similar sources internal to the Airshed which differ from those elsewhere, and shows that the methodology employed here is capable of identifying these differences. The clusters are more distributed across Airsheds for the remaining 5 Airsheds (West Central Airshed Society (WCAS), Fort Air Partnership (FAP), Calgary Regional Airshed Zone (CRAZ), Alberta Capital Airshed Alliance (ACAA), and Peace Airshed Zone Association (PAZA)), though the pattern sometimes identifies commonalities of types of emissions and physical proximity between Airsheds. For example, the continuous monitors at Edmonton Central, Fort Saskatchewan, Edmonton East, Calgary Northwest, Edmonton South, and Red Deer Riverside, and Ross Creek all cluster at 1-R = 0.07 (R=0.93). While these stations are separated greatly in space, they are all located in urban, small town, or mixed urban/industrial settings. The time series of the ambient NO_2 at these locations, and hence their

correlations, will likely reflect urban emissions as a dominant local source of NO_x. The 1-R analysis thus is identifying similar source types separated by large distances, for this cluster.

Referring to Figure 3.11, the first branching shows LICA passive monitor, Primrose, anti-correlating with all the other stations (1-R =1.03), implying that this station may be located in a relatively unique setting (e.g. at a site with a unique set of conditions or sources). Examination of the data shows that Primrose' time series has a large outlier concentration for one of its bimonthly values; the analysis has thus identified Primrose as unique; one of the benefits of this analysis is that such issues are flagged as part of the analysis, and subsequent decisions on data quality assurance and control can be made. All of the PAS passive stations form a separate cluster at a relatively low correlation (R=0.08; 1-R = 0.92). However, PAS's single continuous monitor (Crescent Heights) clusters more closely with other continuous monitors, again suggesting that differences in sampling methodology between the continuous and passive samples. PAMZ's Baseline Mountain passive site correlates poorly with the other stations due to its mountaintop location (R=0.13; 1-R=0.87).

The third main branching at correlation level 0.30 (1-R = 0.70) shows three WBEA passive monitors clustering with PAMZ's Parker Ridge passive monitor – probably a reflection that all four stations have relatively low concentrations and higher signal to noise ratio at such concentrations. Except for St. Lina, all LICA passive monitors cluster together at correlation level 0.83 (1-R = 0.17) and WBEA passive monitors split into two clusters: at correlation level 0.44 (1-R = 0.66) a set of passive monitors are clustering with LICA passive monitors and only at correlation level 0.72 (1-R = 0.27). WBEA passive monitors (R2/9920P, JP205/9910P, JP107/9909P, NE11/9915P, NE10/9914P, SM7/9921P, SM8/9917P, BM10/9903P, and WF4/9918P) form a cluster of their own with JP212/9912P, NE7/9916P, JP213/9913P, BM7/9905P and BM11/9904P falling out as individual clusters when correlation level increases (1-R decreases). The fourth main bifurcation occurs at correlation level 0.65 (1-R = 0.55), where PAMZ passive monitors cluster together at correlation level 0.43 (1-R = 0.57). Also located at PAMZ, passive monitors that stand out of as behaving differently than their pairs: Baseline Mountain, forming a single cluster at correlation level 0.03 (1-R =0.97), Parker Ridge at correlation 0.66 (1-R = 0.34) and Bow Summit at correlation level 0.43 (1-R =0.57). The large cluster resulting from the bifurcation at correlation level 0.68, shows high correlation (0.72) between St Lina, and WCAS continuous monitors: Sleeper and Power; at correlation at level 0.62 (1-R = 0.38), two sets of stations cluster: a smaller cluster comprising of three WBEA continuous stations, Shell Muskeg River/1244C, Fort McMurray-Athabasca Valley/1064C, Millennium Mine/1075C, and a AEP continuous monitoring station (Lethbridge/1049C). The larger cluster first drops at correlation level 0.66 (1-R = 0.34) Fort Chipewyan//1071C and is a cluster of its own and at correlation level 0.79 (1-R = 0.29), WBEA passives (AH7/9919P, AH8/9902P, JP101/9906P, JP102/9907P, AH3/9901P, JP210/9911P, Fort McKay-Bertha Ganter/1032P) and PAMZ (Sunchild/9941P) cluster. PAMZ, CRAZ, WCAS and Alberta Capital Airshed Alliance (ACAA) monitors seem to cluster all together, not finding a specific pattern for these Airsheds but all are continuous monitors with exception of Red Deer Riverside located at PAS. We note that the passive and NO_2 monitor at Fort McKay-Bertha Ganter clusters more closely with other passive monitors both in WBEA and other Airsheds than with the collocated Fort McKay-Bertha Ganter continuous NO_2 monitor.

The NO_2 dendrogram using Euclidean distance as the dissimilarity metric (Figure 3.12) shows that almost all the continuous stations, except Anzac and Fort Chipewyan, separate out from the other stations in the

dendrogram at higher levels of dissimilarity. Single stations are dropping from the dendrogram as single clusters in the first four bifurcations and creating a cluster at a level of 16.9 ppb. At lower levels of Euclidean distance, stations from the same Airshed and the same type of monitor tend to cluster. LICA, PAMZ and WBEA passive monitors are clustering at lower levels of Euclidean distance, with several stations from WBEA and PAMZ that cluster within the same Airshed at very low levels (< 2ppb). The clustering using the 1-R metric (Figure 3.11) shows several groups of stations within Airsheds clustering together (similar colours of station names in the figure being part of the same Airshed), while the clustering using the Euclidean metric (Figure 3.12) does not follow Airsheds to the same extent. Given that the 1-R metric analyzes the data by the shape of the time series, while the Euclidean distance analyzes the data according to the magnitudes of the concentrations, the analysis thus suggests that there is a greater degree of similarity with time series shape, than with the magnitude of reported concentrations, within a given Airshed. The data thus suggest that the time variation of concentration within an Airshed is sufficiently unique that the 1-R metric can identify Airsheds based on that time variation, while the typical magnitude of the concentrations differences between stations is a less unique identifier of an Airshed.

The 1-R metric dendrogram for passive and continuous SO_2 observations (Figure 3.13) shows that correlation between stations is lower than for NO_2 (compare Figure 3.11 and Figure 3.13), although some clustering between groups of stations within Airsheds occurs to a certain level, is mostly seen for WBEA, FAP and LICA stations. These differences may be explained by the differing nature of the types of emitting sources for the two chemicals. SO_2 emissions are dominated by industrial stacks (aka “point sources”), while NO_2 emissions are dominated by surface (also known as “area”) sources. Following emission from an industrial point source, the emitted chemicals will rapidly rise to some height above the source, depending on temperature of the emitted gas, the flow rate of the emissions, the gradient of the temperature, and other meteorological factors in the atmosphere directly above the industrial point source. The direction and speed of the wind may also change significantly at different heights above the industrial point source. The direction of and extent of dispersion of pollutants such as SO_2, emitted from industrial point sources, will thus depend critically on the local meteorological conditions, not just at the surface, but in the region above each industrial point source, up to the height where the rising plume has reached neutral buoyancy and stops rising. In contrast, the dispersion of pollutants such as NO_2, largely emitted from surface sources, will depend less on the changes in meteorology in the region above the source. These additional meteorological factors will tend to make the downwind concentrations of SO_2 from major point sources more variable in time than downwind concentrations of NO_2 from surface area sources. Thus, the type of source, the proximity of a monitoring site to the source, the magnitude of the source as well as meteorological conditions, all influence ambient concentration measured at a site. Hence a greater degree of variability (hence lower R values) in the clustering of monitoring sites is expected for SO_2 than for NO_2. The first branching of the dendrogram shown in Figure 3.13 is an example of the variability expected, dividing two large sets of stations that are anti-correlating (R= - 0.11, 1-R = 1.11), with majority of LICA and PAS stations clustering in one side and WBEA stations in another.

The colour-coding of the stations according to Airshed in Figure 3.13 shows that some clusters occur within airsheds, at relatively low values of 1-R (highly correlated records); examples are identified by enclosed boxes in the figure. Fort Air Partnership (FAP, green) stations 15, 22, 28, 03, 40 and 46 are all within one cluster with a 1-R value of 0.25 (R=0.75), FAP stations 02, 20, 21, and 47 cluster for 1-R of 0.2

(R=0.8), and thirteen LICA stations (black) cluster at 1-R of 0.25. There are several clusters of WBEA stations at lower levels of correlation (e.g. five stations at the far right of Figure 3.13 clustering at 1-R of 0.37, next five WBEA stations from the right cluster at 1-R of 0.6, WBEA stations SM8, Fort Chipewyan, BM7, NE10, and AH3 cluster at 1-R=0.8. Much of the remaining clustering shows 1-R similarities which sometimes link stations that are widely separated in space, such as WBEA station Fort McMurray-Patricia McInnes clustering at 1-R of 0.80 with stations such as Tomahawk and Violet Grove in the WCAS airshed.

The SO_2 dendrogram using the Euclidean distance metric is shown in Figure 3.14. Clusters for some monitoring sites within an Airshed occurs, leading to the possibility that some groups of stations may be associated with clusters which have high correlations and low Euclidean distances, hence may be more redundant from the standpoint of both metrics. This will be examined in more detail in Section 4. The magnitudes of the Euclidean distances are uniformly small for much of Figure 3.14, indicating the presence of a large number of stations with similar concentration records; these may represent the influence low concentration values (~1 ppb). This is noticeable in particular for some of the FAP stations mentioned in the context of the previous figure, many of which are clustered at Euclidean distances of 1 to 3 ppbv.

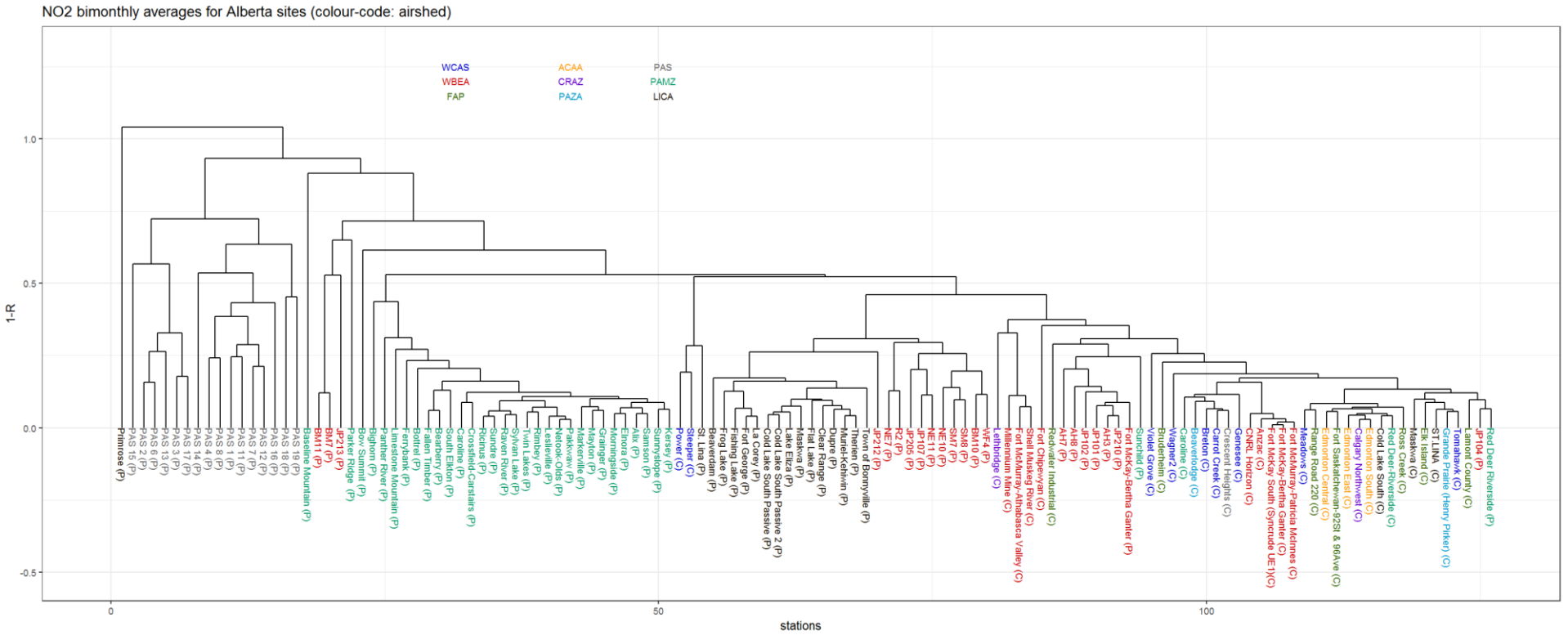


Figure 3.11 Dendrogram analysis for passive and continuous bimonthly NO_2 averages considering 1-R as the metric to compute the dissimilarity matrix, for West Central Airshed Society (WCAS), Wood Buffalo Environmental Association (WBEA), Fort Air Partnership (FAP), Alberta Capital Airshed Alliance (ACAA), Calgary Regional Airshed Zone (CRAZ), Peace Airshed Zone Association (PAZA), Palliser Airshed Society (PAS), Parkland Airshed Management Zone (PAMZ) and Lakeland Industrial Community Association (LICA). Stations are colour-codded according to Airshed. Station names which are continuous end in a "(C)", and stations which are passive end in a "(P)".

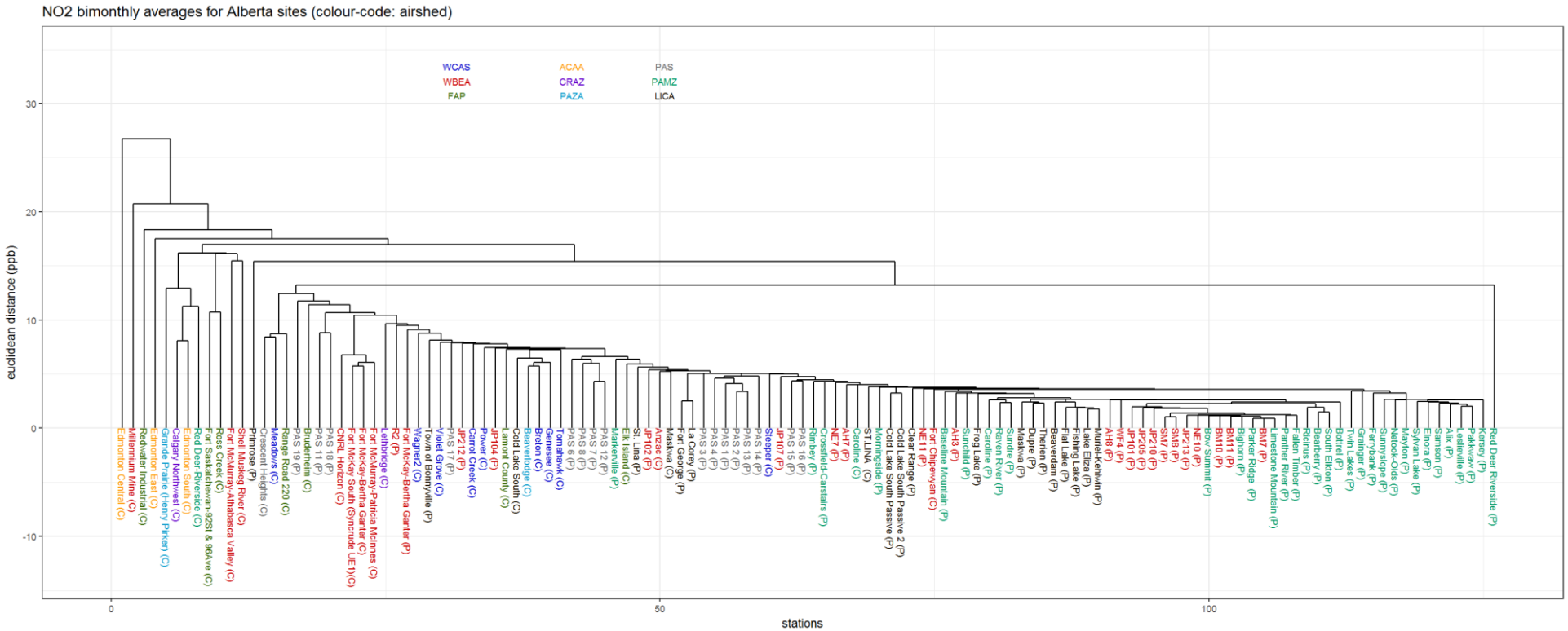


Figure 3.12 Dendrogram analysis for passive and continuous bimonthly NO_2 averages considering Euclidean distance (ppb) as the metric to compute the dissimilarity matrix, for West Central Airshed Society (WCAS), Wood Buffalo Environmental Association (WBEA), Fort Air Partnership (FAP), Alberta Capital Airshed Alliance (ACAA), Calgary Regional Airshed Zone (CRAZ), Peace Airshed Zone Association (PAZA), Palliser Airshed Society (PAS), Parkland Airshed Management Zone (PAMZ) and Lakeland Industrial Community Association (LICA). Stations are colour-codded according to Airshed. Station names which are continuous end in a "(C)", and stations which are passive end in a "(P)".

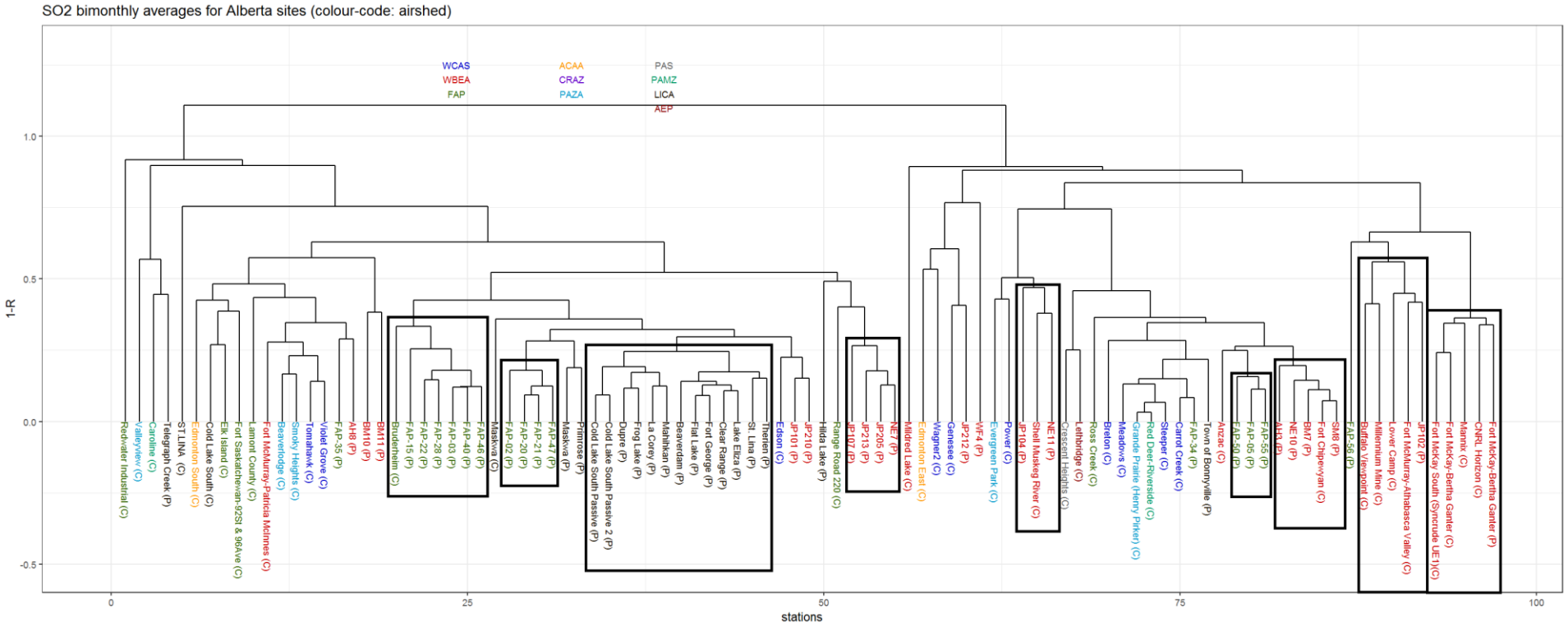


Figure 3.13 Dendrogram analysis for passive and continuous bimonthly SO_2 averages using 1-R as the metric to compute the dissimilarity matrix, for West Central Airshed Society (WCAS), Wood Buffalo Environmental Association (WBEA), Fort Air Partnership (FAP), Alberta Capital Airshed Alliance (ACAA), Calgary Regional Airshed Zone (CRAZ), Peace Airshed Zone Association (PAZA), Palliser Airshed Society (PAS), Parkland Airshed Management Zone (PAMZ) and Lakeland Industrial Community Association (LICA). Stations are colour-codded according to Airshed. Station names which are continuous end in a “(C)”, and stations which are passive end in a “(P)”. The back boxes identify stations clustering within the same Airshed at low dissimilarity levels.

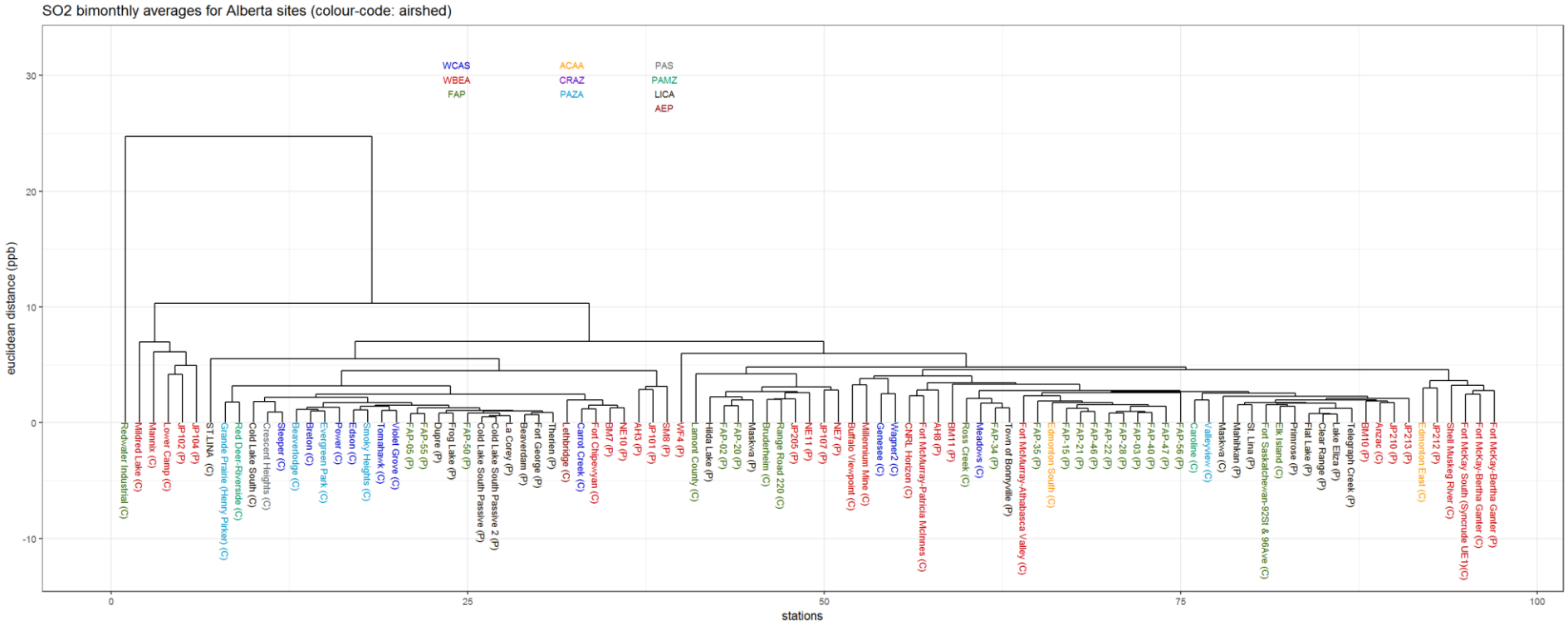


Figure 3.14 Dendrogram analysis for passive and continuous bimonthly SO_2 averages considering Euclidean distance (ppb) as the metric to compute the dissimilarity matrix for West Central Airshed Society (WCAS), Wood Buffalo Environmental Association (WBEA), Fort Air Partnership (FAP), Alberta Capital Airshed Alliance (ACAA), Calgary Regional Airshed Zone (CRAZ), Peace Airshed Zone Association (PAZA), Palliser Airshed Society (PAS), Parkland Airshed Management Zone (PAMZ) and Lakeland Industrial Community Association (LICA). Stations are colour-coded according to Airshed. Station names which are continuous end in a "(C)", and stations which are passive end in a "(P)".

3.4 Associativity Analysis for Alberta Province: Continuous Hourly Observations

The hierarchal clustering method as described in Section 2 was applied to analyze hourly observations for O_3, NO, NO_2, NO_x, $PM_{2.5}$, SO_2, NMHC, THC, TRS and CH_4 for a period between August 1, 2013 and July 31, 2014 with overlaps before and after this period to accommodate the KZ filtering of the central period of one year. Data selection and QA/QC is described in Section 2.2.2. The time interval was chosen for two main reasons:

(1) The first two months of this period correspond to the time interval during which a joint ECCC/AEP monitoring intensive took place in the oil sands region, allowing for possible comparisons with monitoring intensive instrumentation in other studies.

(2) The period corresponds to a full year simulation carried out by the ECCC air pollution model GEM-MACH (Makar et al, 2015(a,b), Makar et al 2017, Moran et al, 2010) – this simulation will be used in later phases of the network analysis project, and has been used here (see Section 4.3) to examine the issue of sampling errors on clustering and redundancy calculations.

For each of the chemical species, continuous hourly data was filtered to remove the time scales smaller than a day, a week, and a month, applying the KZ-filtering, as described in Section 1.2. The results of subsequent clustering analyses using the 1-R and Euclidean distance metrics are shown in pairs of figures which follow; the 1-R figure first, followed by the Euclidean distance figure. As before, the station names are colour-coded by the respective Airshed.

Our analysis begins with O_3 (Figure 3.15 and Figure 3.16). Figure 3.15 (1-R dissimilarities) show that the stations cluster largely according to Airshed, with the exceptions of WCAS' Steeper/1055 and Genesse/1057 stations, and PAMZ's two stations at Red Deer-Riverside/1142 and Caroline/1092. Steeper station is located at relatively higher elevation (1400m asl) and thus samples air more influenced by the middle to upper Troposphere than the other stations within the WCAS, while Genesse is relatively close to a coal-fired power-plant and thus may be expected to be more impacted by NO_x fumigation and ozone titration events from powerplant plumes than other stations in the WCAS region. We note that Red-Deer-Riverside samples urban air while Caroline is located in a rural location in the foothills at 1100m above sea level altitude, hence may be expected to cluster more closely with the relatively alpine Steeper station than Red-Deer-Riverside – as was seen in the analysis. The WBEA continuous O3 stations, as well as the remaining stations in WCAS, PAZA, CRAZ, FAP, and ACAA tend to cluster within Airsheds rather than across Airsheds for the analyses with hourly, daily, and most of the weekly and shorter periods removed dendrograms (Figure 3.15 (a),(b),(c), respectively). This shows that the time variation of ozone is more affected by local, rather than regional influences, at time scales of less than a month. When time scales less than one month are removed (Figure 3.15 (d)), this within-airshed clustering starts to be lost, with the stations clustering across Airsheds to a greater degree. For example, at time scales of greater than one day, the Lethbridge and Medicine Hat stations form a cluster, and remain clustered thereafter – these stations are both in Southern Alberta, downwind of the Rockies, and might be expected to cluster on regional transport time scales. On timescales longer than monthly, the larger urban stations form a high correlation cluster (Figure 3.15 (d), right-hand side), indicating that the ozone in these

locations even at long time scales is affected more by local conditions and a common pattern of urban emissions than regional influences. Four WBEA stations remain clustered throughout the period, indicating that their ozone levels are affected by common sources (Fort Mckay-Bertha Ganter/1032, Fort McMurray-Particia McInnes/1070, Fort McKay South/1076, and Fort McMurray-Athabasca Valley/1064.

The NO_2 dendrograms considering 1-R as a metric to compute the dissimilarity matrix (Figure 3.15) generally shows clustering between stations within the same Airshed. An increase in the clustering of stations between different Airsheds can be seen as the shorter time scales are progressively removed, but the clustering within some Airsheds (WBEA and FAP) seems less affected by the filtering of shorter time variability, suggesting that the observations at these stations within these Airsheds are more similar across multiple time scales than they are to the observations at other airsheds. Correlation levels between stations improve as KZ filtering is applied and shorter time variabilities are removed

The NO_2 1-R dendrogram (Figure 3.17) shows 1-R clustering by Airshed for WBEA, FAP, ACAA and PAZA Airsheds for hourly data. Like O_3, clustering by Airshed becomes less prominent as shorter time scales are removed, indicating that much of the short term variation in NO2 is due to local sources. The exceptions are stations within WBEA, which remain clustered even at monthly and greater timescales (Figure 3.17 (d)). This indicates NO_2 concentrations measured at WBEA stations have notably different temporal variability from stations located elsewhere, and thus are dissimilar to all other stations in the dataset, even at timescales of greater than a month. An alternative way of putting this: the time series of concentrations observed at these stations are highly similar within the airshed, and highly dissimilar to stations outside of the airshed, at all time scales, for the metrics used here. This in turn suggests that there are aspects of the combination of local emissions and meteorology that ensures that the WBEA stations are “unique”, i.e. more similar within the airshed than to stations outside of the airshed. At the same time, the Euclidean distance magnitudes observed at the WBEA stations vary with timescale (Figure 3.18), with the Fort McKay-Bertha Ganter/1032C, Fort McMurray Patricia McInnes/1070C, Fort McKay South/1076C and CNRL Horizon/1226C stations forming one cluster, and Shell Muskeg River/1244C , Millennium Mine/1075C, and Fort McMurray-Athabasca Valley/1064C stations forming another cluster at time scales greater than one week, and greater than one month (Figure 3.18(c,d)). As with O_3, the NO_2 dendrograms show increasing correlation levels (1-R, Figure 3.17), and Euclidean distances decrease (Figure 3.18), as progressively larger timescales are filtered from the data, indicating a tendency for concentrations with smaller time scale variability removed to be very similar, and much of the short term variability in measured concentrations to be likely due to local sources.

NO dendrograms for the two metrics, 1-R and Euclidean distance, are shown in Figure 3.19 and Figure 3.20, respectively. There are some interesting differences between the NO dendrograms and the NO_2 dendrograms described above, probably driven by NO being a better indicator of very fresh emissions of NO_x, and NO_2 being a better indicator of transport and downwind chemistry. As before, 1-R dendrograms tend to cluster by Airshed when shorter timescales variability in the data are retained (Figure 3.19 (a), and only the WBEA stations remain clustered largely as a group at timescales longer than a month (Figure 3.19 (d)). However, for those longer timescales, NO is also seen to be 1-R clustering by source type and location, with many of the urban stations forming a single cluster to the right of Figure 3.19(d)). Figure 3.19 also shows clustering of similar stations influenced by source types at varying levels of correlation – e.g. Genesee and Wagner2 stations, both influenced by coal-fired power plants, remain clustered at all

time-filtering levels in Figure 3.19. In contrast to NO2, Euclidean distances for NO (Figure 3.20) tend to be less associated with specific Airsheds, and more with specific sources, and relatively little information for clustering is sometimes left once monthly and shorter timescales have been removed (that is, clustering between different Airsheds occurs). Stations which are disconnected in terms of sources are seen to Euclidean distance cluster even at hourly time scales (Figure 3.20 (a), e.g. St. Lina and Steeper stations both have a relatively low Euclidean distance). This is more a measure of both stations having sufficiently low concentrations that they are rated as highly similar using a Euclidean distance metric, than an association based on the influence of local sources. This issue will be discussed further in Section 4.5.

Concentrations of NO_x, and consequently the clustering for the 1-R and Euclidean metrics (Figure 3.21 and Figure 3.22)) are dominated by the NO_2 component, hence tend to follow the NO_2 behaviour more closely than NO (e.g. WBEA stations remain in 1-R clusters at all time scales, a general tendency to lose within-Airshed 1-R clustering outside of WBEA as successively longer timescales are removed, and substantial decreases in 1-R and Euclidean distance as longer timescales are removed, with the latter indicating most of the signal resides in the shorter time scales. The NO_x values remain tightly clustered for WBEA at time scales longer than monthly (Figure 3.21 (d)) suggesting significant local source signal influence remains even at monthly time scales for this region, while the other monitoring association Airsheds show broader scale (or low concentration) influences at the longer time scales.

The SO_2 dendrograms for 1-R differ from the other species examined thus far in terms of the lower level of R values (Figure 3.23); a greater degree of dissimilarity for SO_2 may be seen than for NO_X or O_3 for this metric. This is due to the nature of the sources of SO_2, which are almost exclusively from industrial point sources. The direction and concentration of the plumes is thus highly variable in time, and concentrations from the same source may not correlate as well between two downwind stations, if they are not directly in line downwind. Despite this, the SO_2 1-R dendrograms show the WBEA stations as a single albeit low R level cluster for the first two time levels (Figure 3.23 (a, b)), and most WBEA stations fall into a single cluster even at timescales of greater than one week (Figure 3.23 (c)), with clustering within smaller sets of WBEA stations thereafter. The 1-R dendrograms for concentrations with shorter timescale variability removed are clearly showing clusters forming between widely separated locations. More than many of the other species considered in this report, SO_2 concentrations contain spike-like short time-scale variations, and the removal of the short time scales results in notably lowering of the residual concentrations (an observation also noted to a lesser degree for NO_x). 1-R clustering for monthly and longer timescales (Figure 3.23 (d)) probably reflects the extent to which the filtered time series are all close to zero, as opposed to true relationships across different locations, for SO_2. Euclidean distances also show clustering by Airshed being maintained to timescales of greater than a day, but as successively longer timescales are removed, the clusters tend to be across Airsheds (with the exception of some subgroups of WBEA stations). This also is an indication that most of the concentration magnitude signal resides in the shorter time scales.

$PM_{2.5}$ dendrograms for 1-R and Euclidean distances (Figure 3.24 and Figure 3.25) maintain the WBEA and WCAS stations (aside from the Steeper, located on a ridge-top at roughly 1410m elevation and further into the foothills compared to the other WCAS stations) as being independent clusters at all time scales examined. Both the shape of the time series and the concentration magnitudes are thus dominated by short term variability likely due to local sources and conditions within these Airsheds. For

the WBEA stations, a significant source of local $PM_{2.5}$ is likely the fugitive dust from the oil sands open pit facilities and processing, which has been found to contain a relatively small fraction of secondary species in surface-based observations (Wang et al, 2015). The WCAS stations aside from Steeper are located in a region influenced by open-pit mining of coal (a primary particulate source) and coal-fired power-plants (potentially a source of secondary PM). Which of these two sources dominates the signal observed in the analysis can't be determined in the absence of speciation information, though the similarity across stations within each Airshed shows that they are being influenced by similar sources, suggesting that within these Airsheds there are relatively unique processes determining the average concentration of $PM_{2.5}$. The Euclidean distances also cluster for 4 of the WCAS stations beyond filtering times of 1 month (Figure 3.25 (d)), indicating a single regional source for $PM_{2.5}$ at longer time scales. At shorter timescales, the Euclidean distance (Figure 3.25 (a), b) clusters almost exclusively by Airshed (exceptions Steeper station and Hinton Station), suggesting within-Airshed emissions and atmospheric control the origin of $PM_{2.5}$ at these timescales.

CH_4 dendrograms show 1-R clustering for most Airsheds represented at hourly timescales (Figure 3.27(a)), for WBEA stations up to timescales greater than 1 day (Figure 3.27 (b)), and all Airshed-related clustering is removed once monthly and longer timescales are removed (Figure 3.27 (d)). A similar pattern can be seen with the Euclidean distances (Figure 3.28), with within-Airshed clustering being retained for some pairs when weekly and shorter timescales are removed (Figure 3.28(c)) except for one WBEA pair which still clusters for timescales longer than monthly (Figure 3.28 (d))). The implication is that on an hourly basis, methane concentration time series shape and magnitude are being controlled by local sources, but on longer timescales, the regional background levels start to dominate, with the exception of Fort McKay-Bertha Ganter and Fort McMurray-Athabasca Valley in the WBEA region. These sites are likely impacted by notable local sources.

THC dendrograms have relatively low initial correlations (Figure 3.29(a)), indicating greater variability between sources. 1-R clustering within Airsheds is maintained up to timescales greater than daily (Figure 3.29(b)) and then begins to break down, though sub-groups of WBEA stations remain clustered with each other up to timescales greater than monthly (Figure 3.29(d)). THC Euclidean distances, however, do not follow Airsheds even at hourly timescales (Figure 3.30). The magnitude of THC concentrations measured at monitoring stations is thus more dissimilar despite similarities in the variation of concentration over time. One possible reason for the high within-Airshed correlation similarities and lower similarities based on Euclidean distance may be proximity from the sources: two stations along the direction of the prevailing wind from an upwind emissions source will be highly correlated, but their Euclidean distances may also be high due to the additional dispersion between the two stations. Euclidean distance may therefore sometimes cluster more highly across Airsheds than within a given airshed.

NMHC dendrograms have very low hourly correlations, though they follow Airsheds up to timescales of greater than daily (Figure 3.31(b)), and some two-member clusters within Airsheds are maintained to monthly and greater timescales (Figure 3.31(d)). Euclidean distance clustering even at hourly scales (Figure 3.32) fails to follow Airsheds, with a similar explanation for these results as for THC.

TRS observations are only available from three Airsheds (PAMZ, one station; PAZA, 4 stations; WBEA, 8 stations). The WBEA and PAZA stations maintain 1-R clustering up to timescales greater than daily being removed (Figure 3.33(a),(b)), with separate clusters forming at longer timescales (and curiously reforming

as a single cluster for monthly and longer timescales for all WBEA stations aside from Millennium Mine, see Figure 3.33(d)). Both 1-R and Euclidean distances (Figure 3.34) show Hinton station as being very different from the other stations, probably due to the relatively unique sources of TRS (in type, magnitude and frequency of events) near Hinton (for example, the local pulp and paper mill, as opposed to the oil and gas industry sources at the other sites). The lack of within-Airshed Euclidean distance clustering at even hourly timescales for TRS, similar to THC, NMHC, and to a lesser extent CH_4, again suggests a rapid drop in co-varying concentrations with distance from sources.

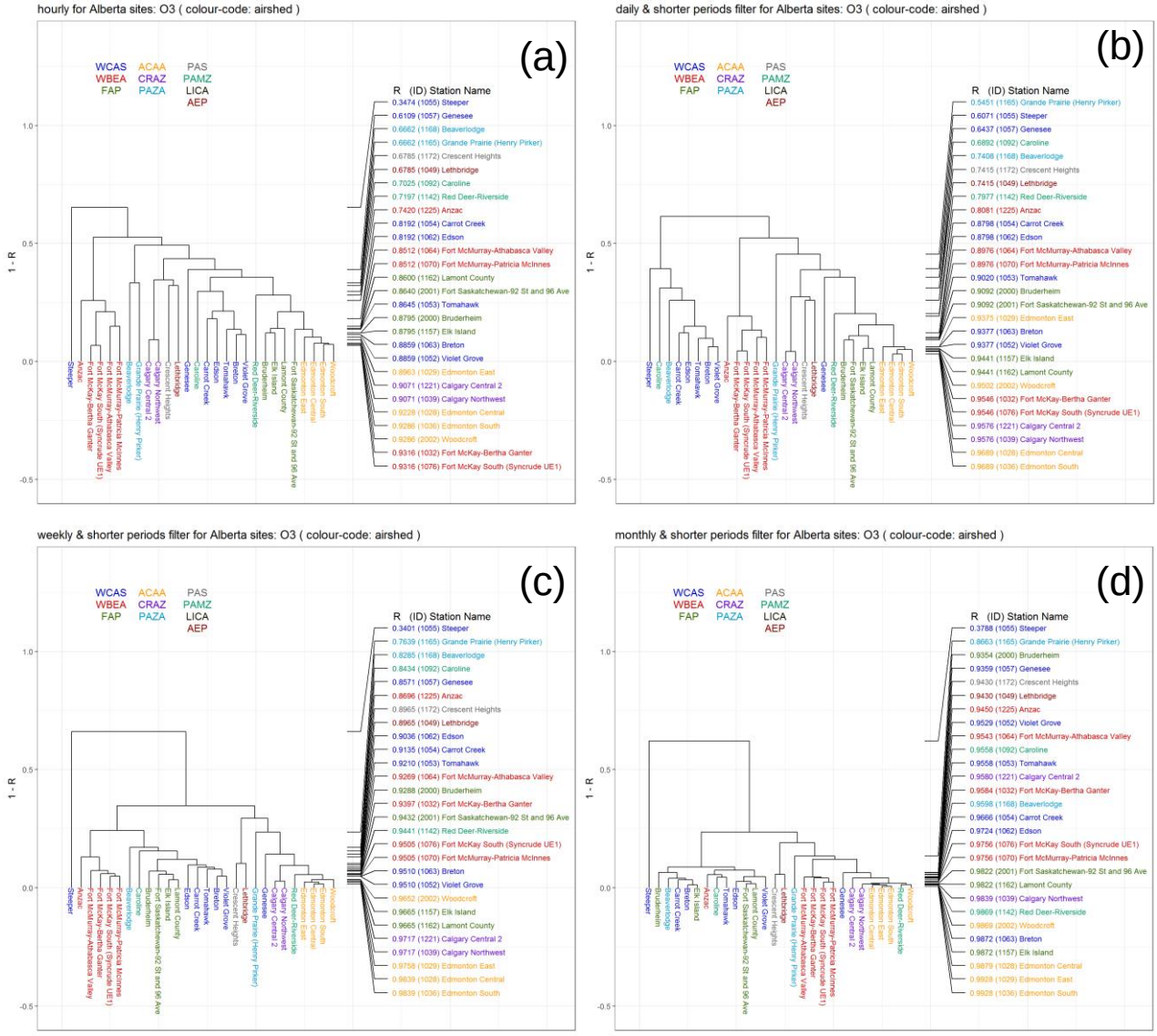


Figure 3.15 Continuous O3 1-R dendrogram analysis, (a) hourly and filtered ((b) daily, (c) weekly and (d) monthly scales removed). Airshed names: WCAS: West Central Airshed Society, WBEA: Wood Buffalo Environmental Association, FAP: Fort Air Partnership, ACAA: Alberta Capital Airshed Alliance, CRAZ: Calgary Regional Airshed Zone, PAZA: Peace Airshed Zone Association, PAS: Palliser Airshed Society, PAMZ: Parkland Airshed Management Zone, LICA: Lakeland Industrial Community Association (LICA). Stations are colour-coded according to Airshed.

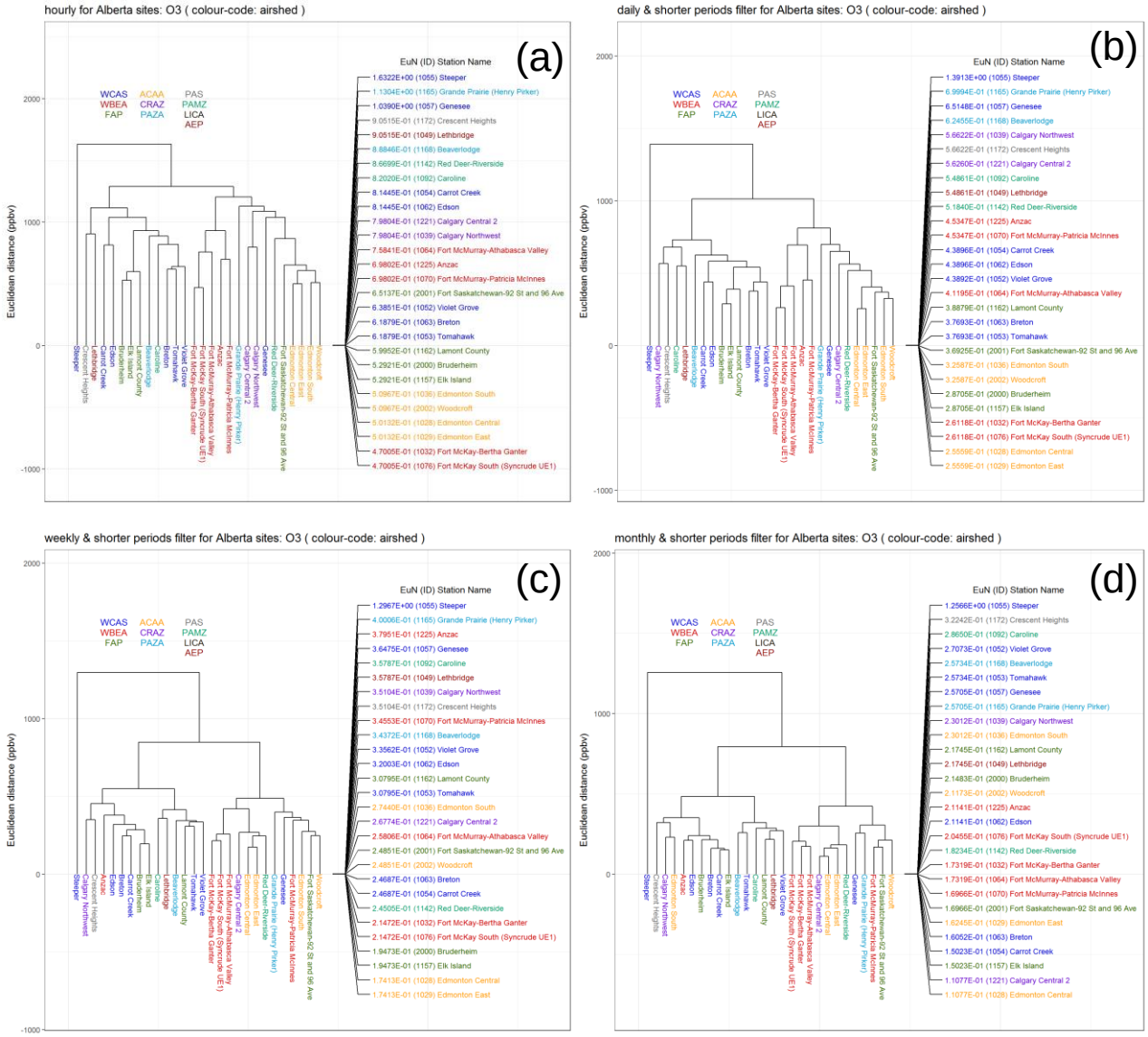


Figure 3.16 Continuous O_3 Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

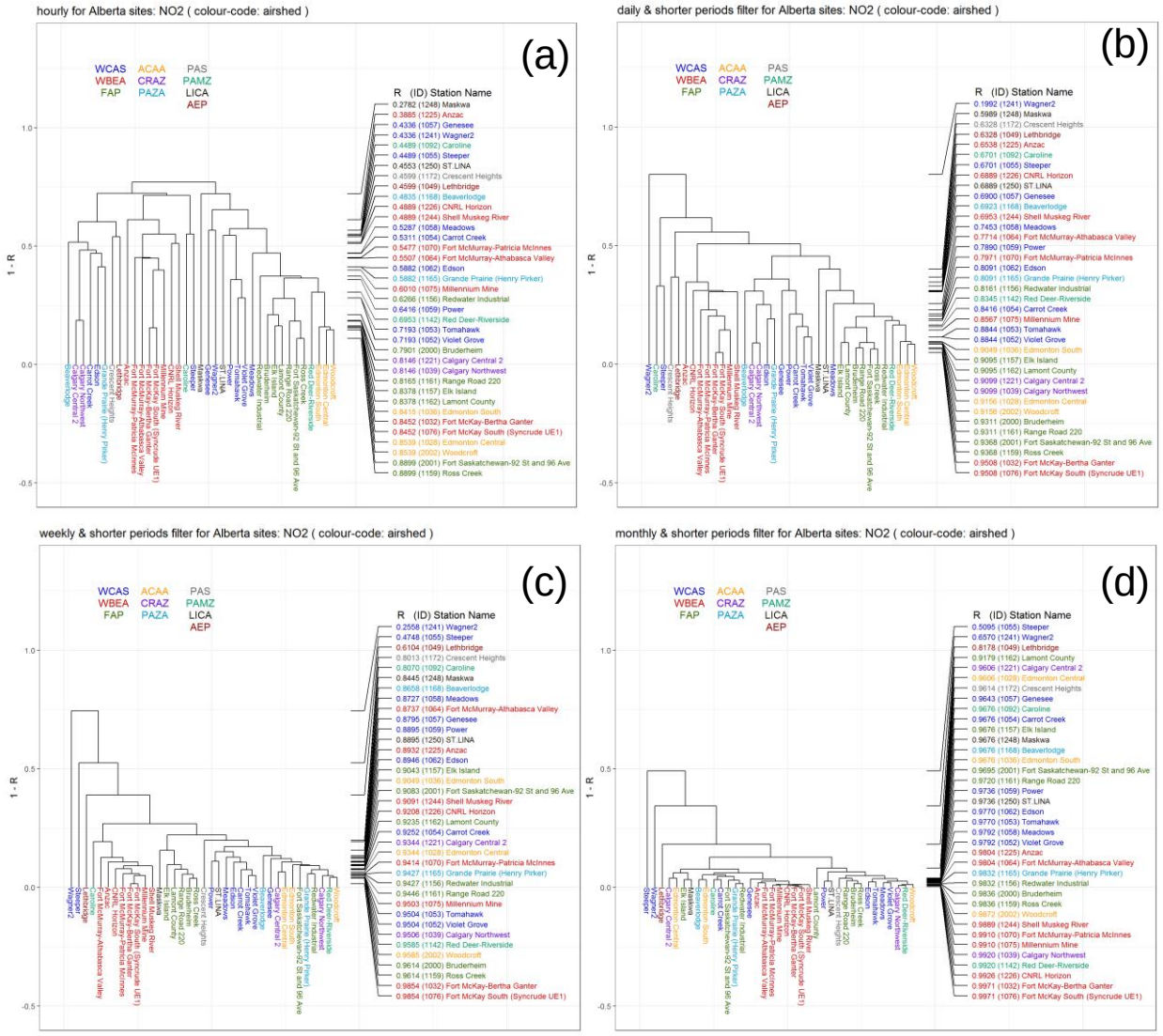


Figure 3.17 Continuous NO_2 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

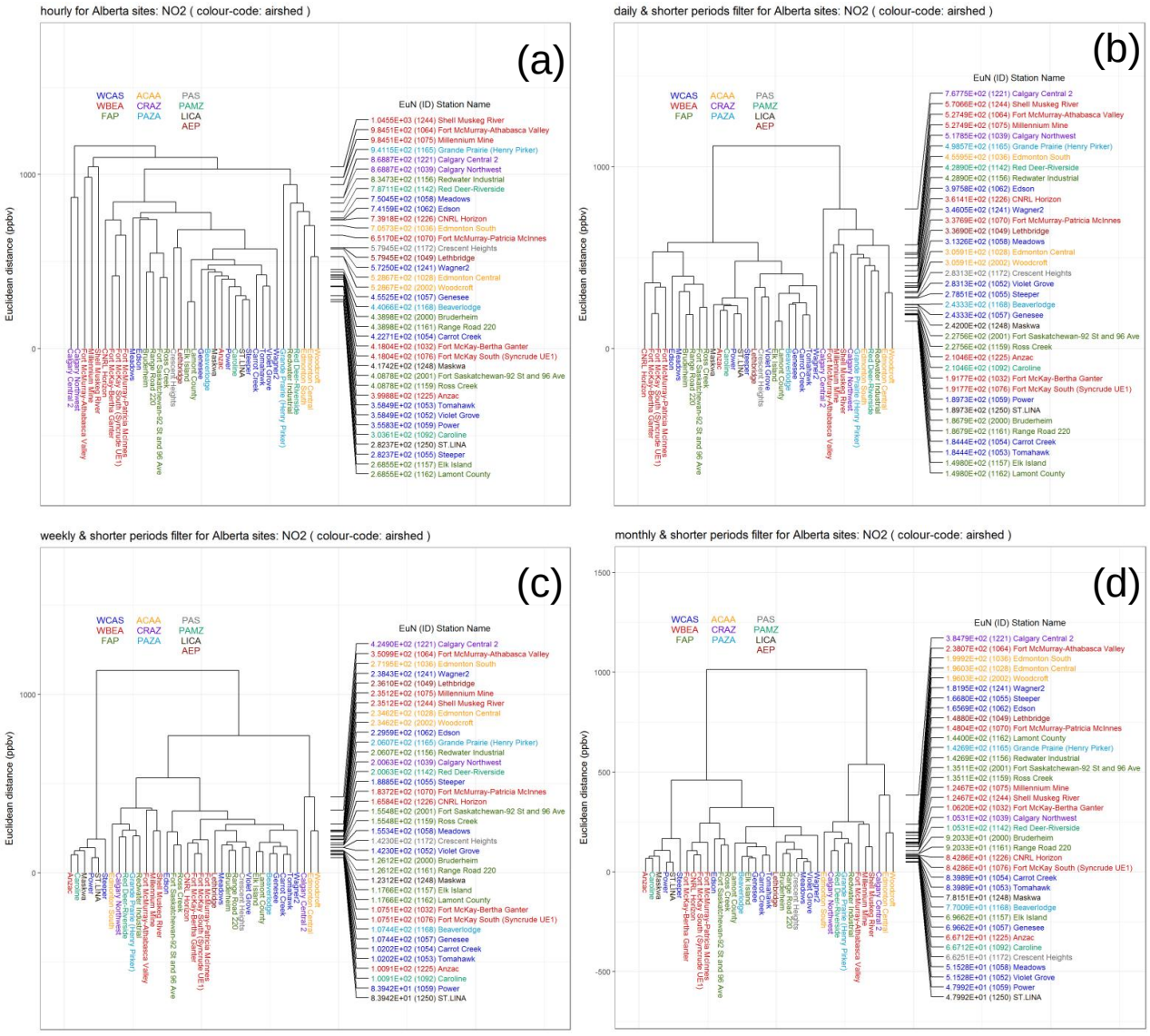


Figure 3.18 Continuous NO_2 Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

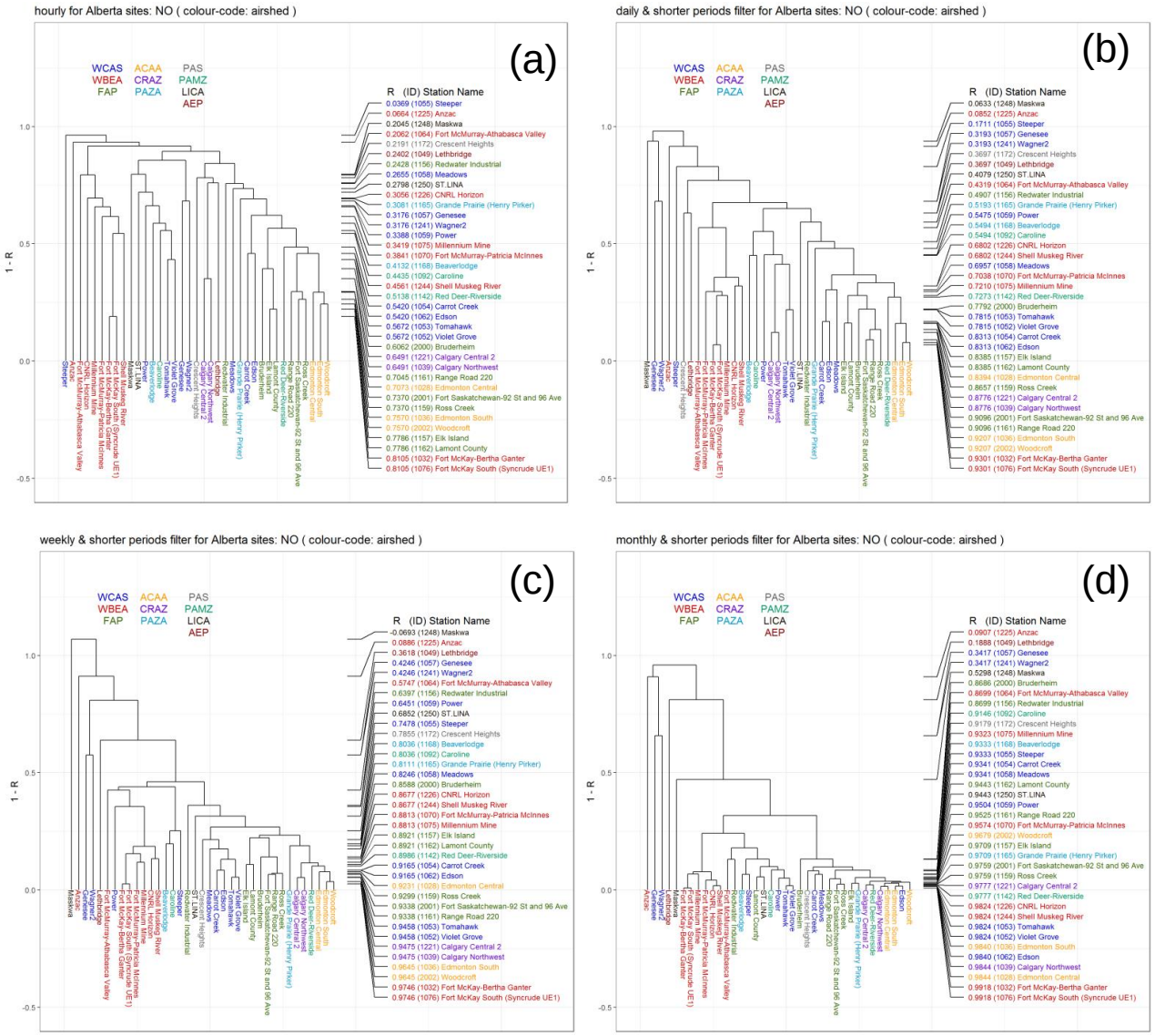


Figure 3.19 Continuous NO 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

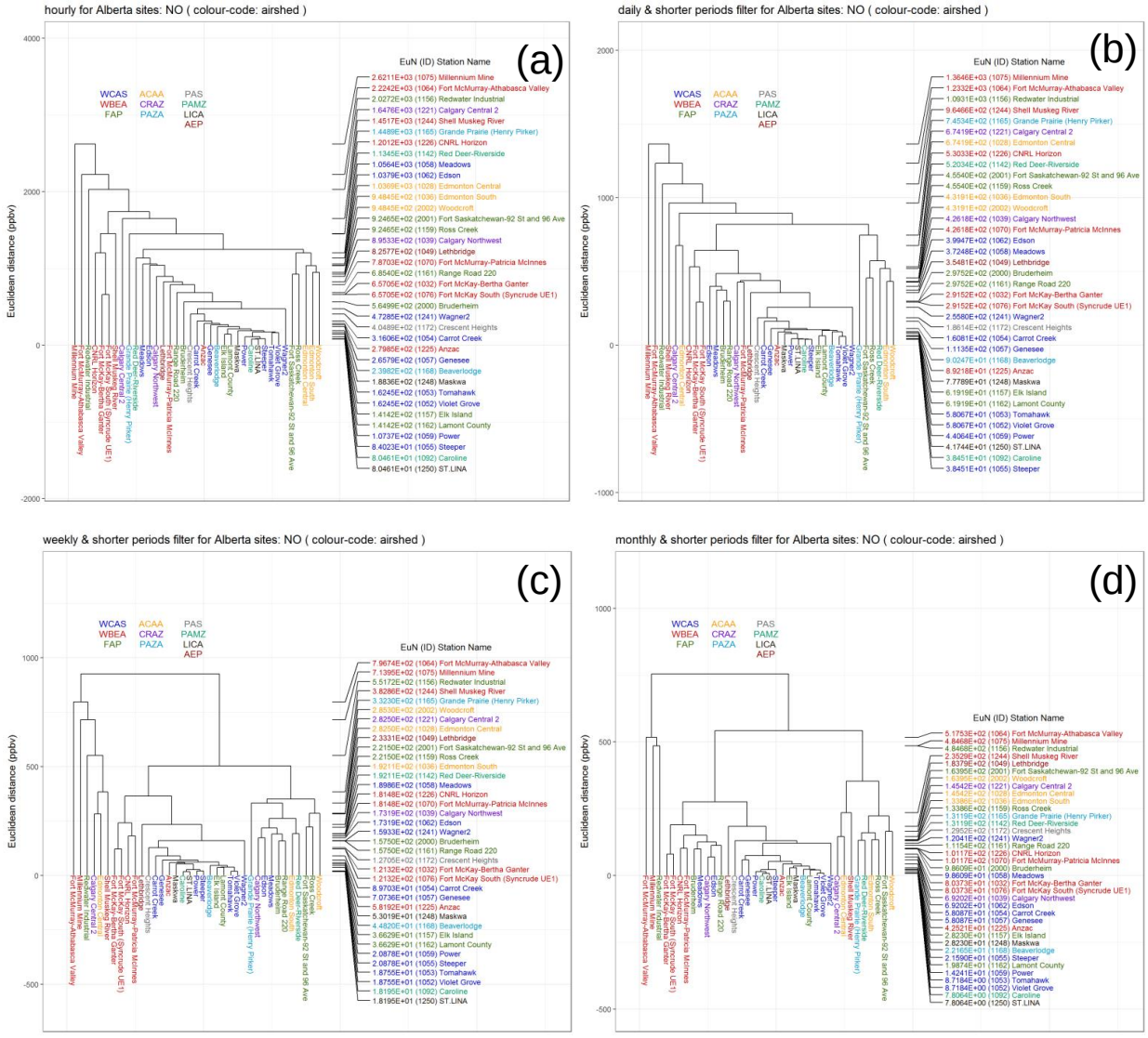


Figure 3.20 Continuous NO Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

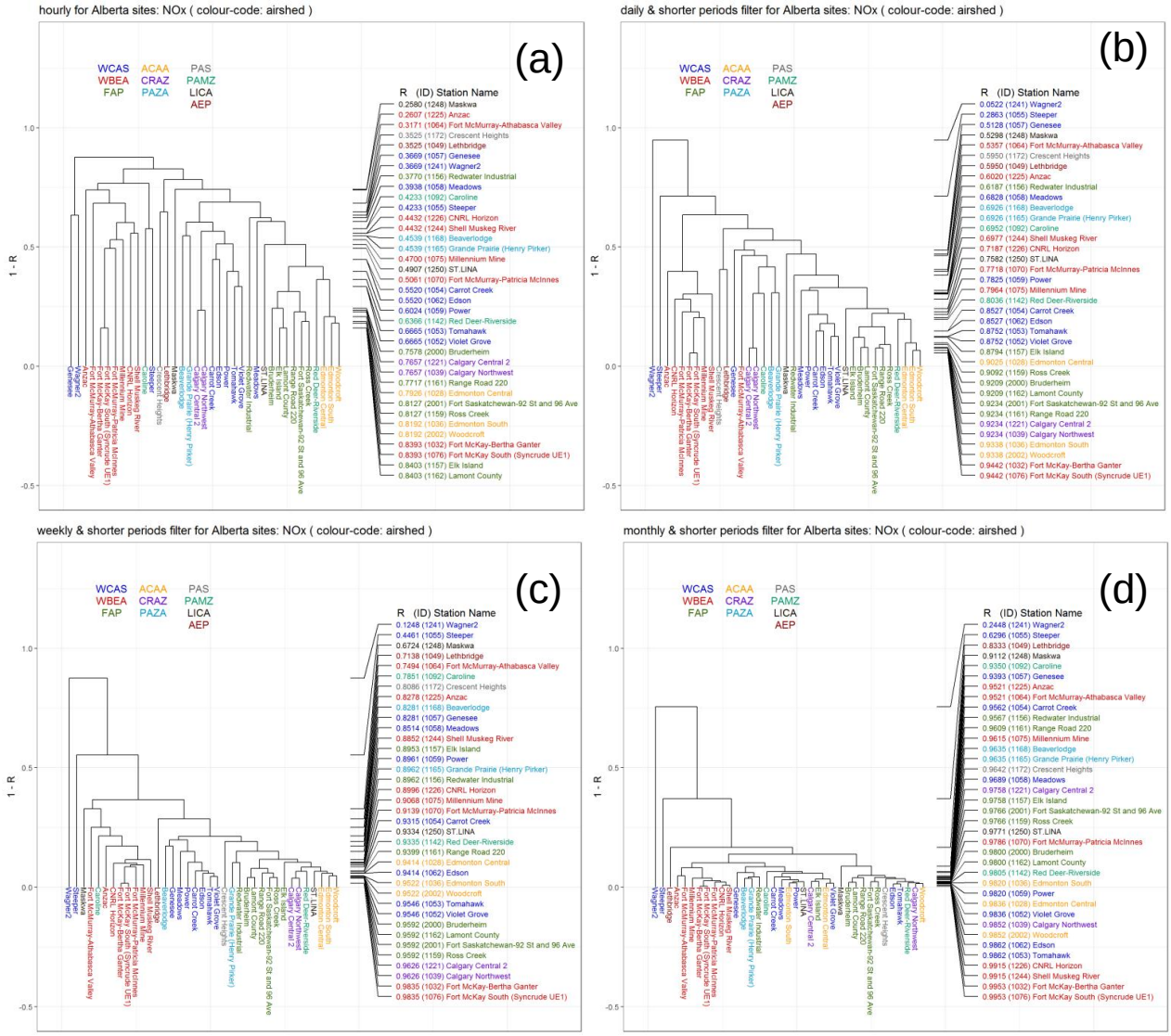


Figure 3.21 Continuous NO_x 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

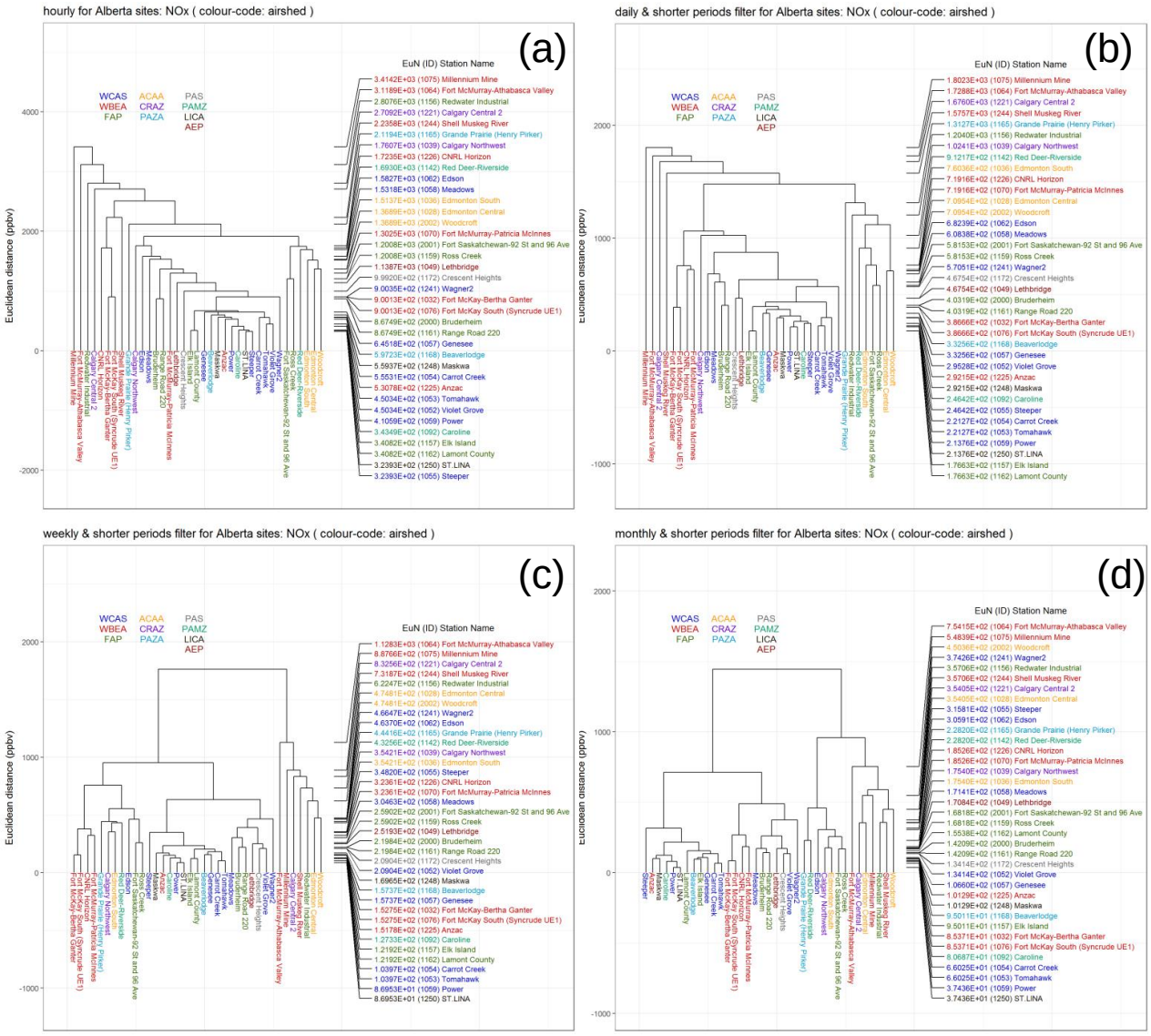


Figure 3.22 Continuous NO_x Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

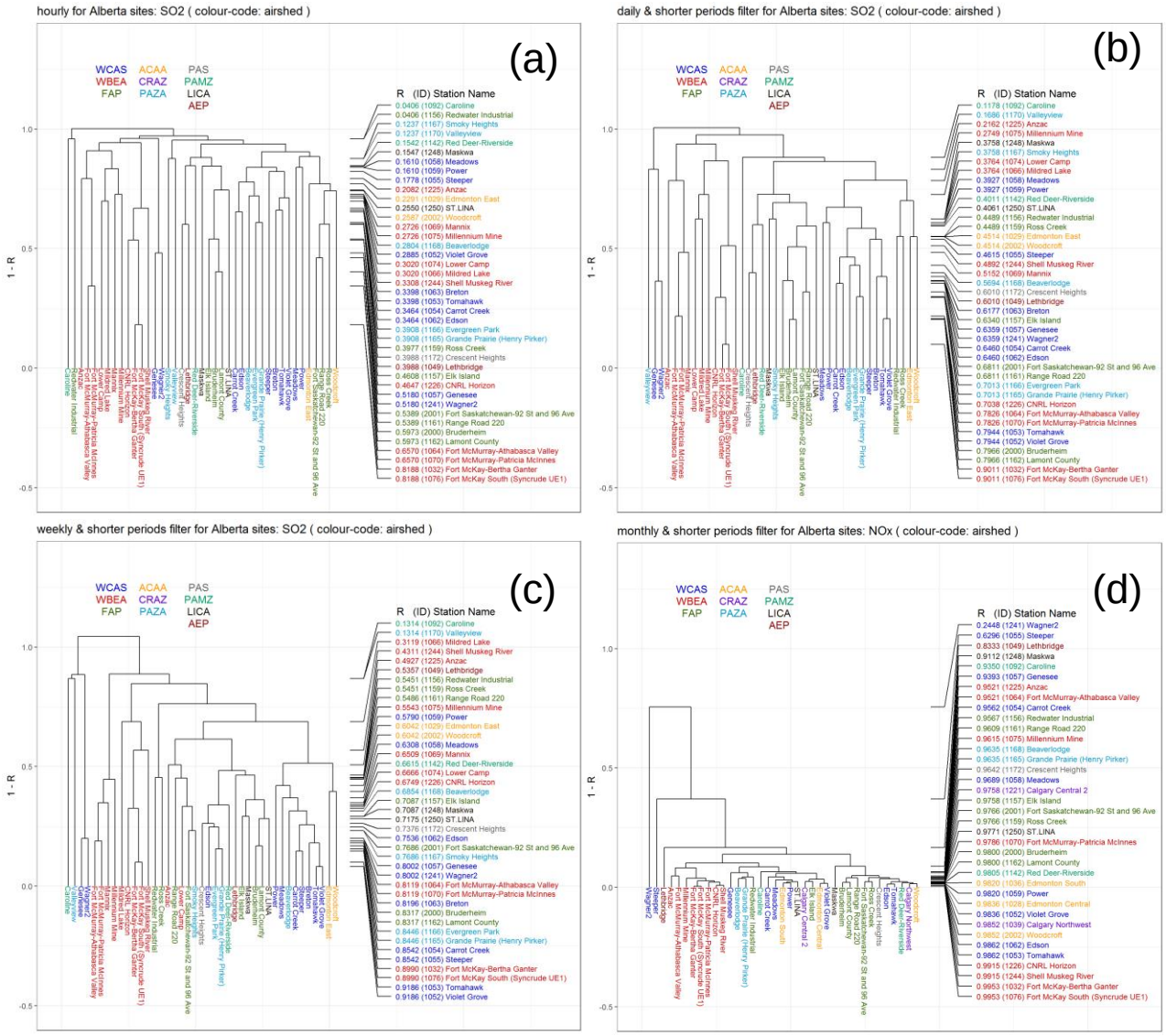


Figure 3.23 Continuous SO_2 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

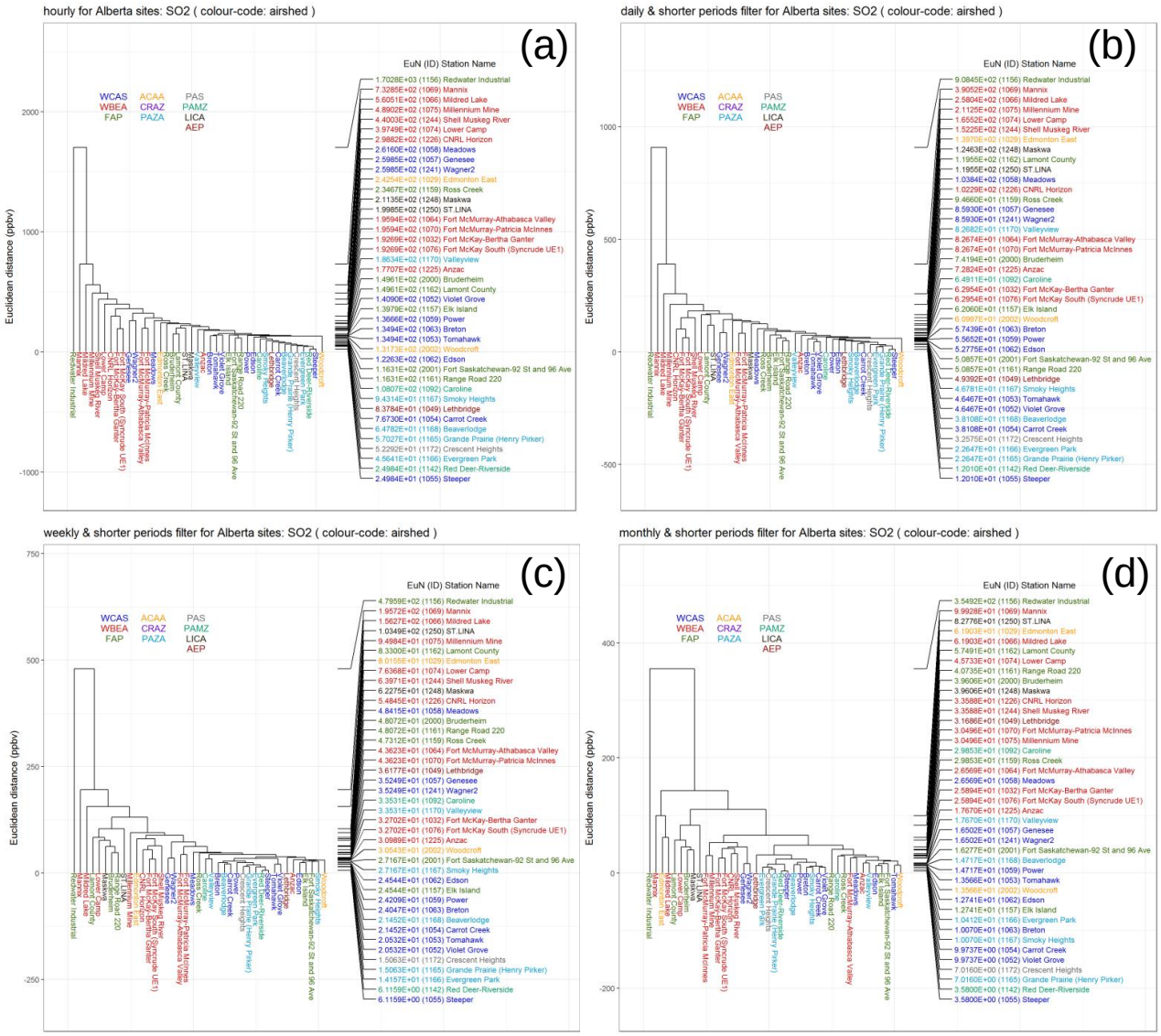


Figure 3.24 Continuous SO_2 Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15. Note that the vertical scale changes between the panels of this figure.

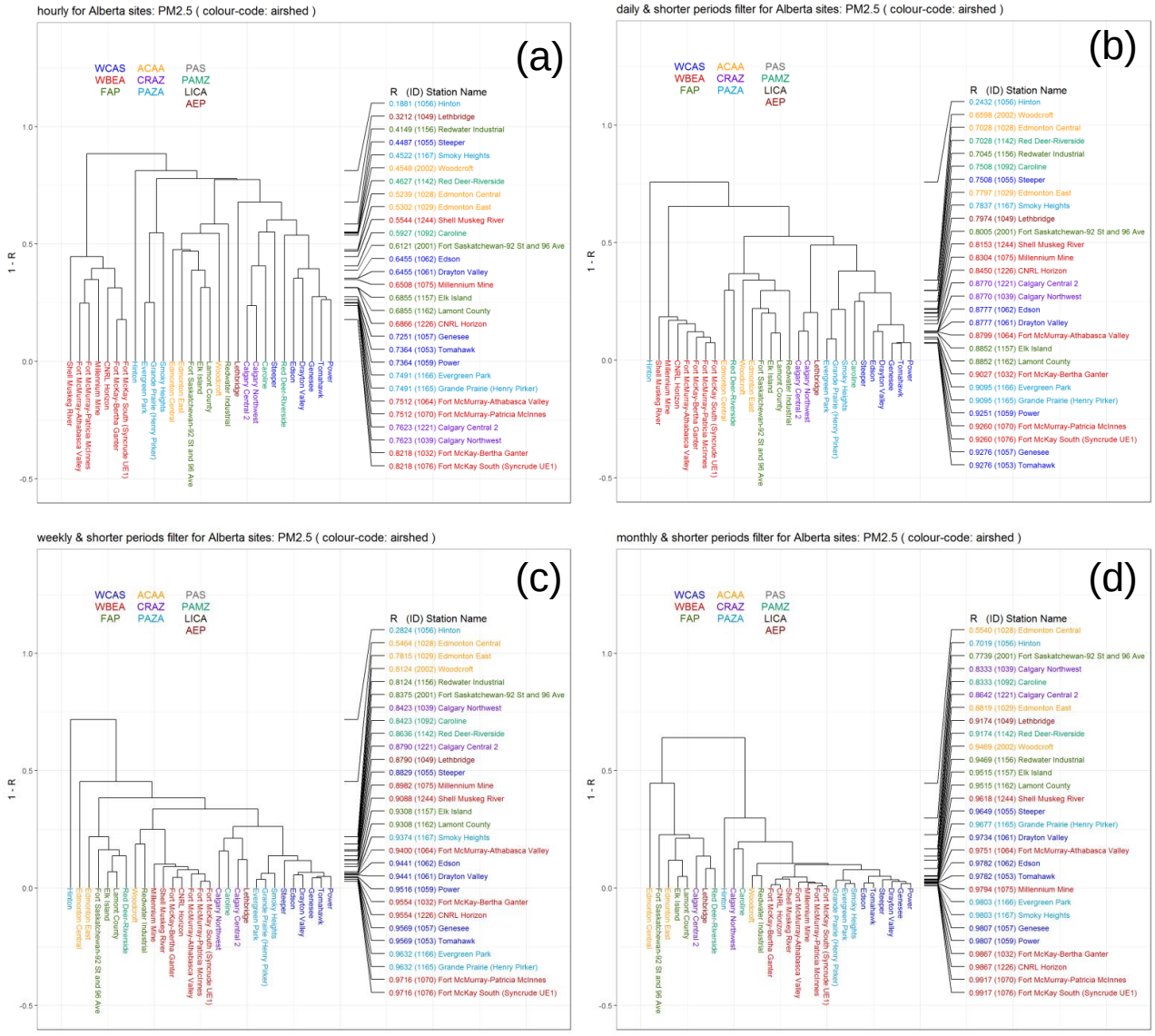


Figure 3.25 Continuous $PM_{2.5}$ 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

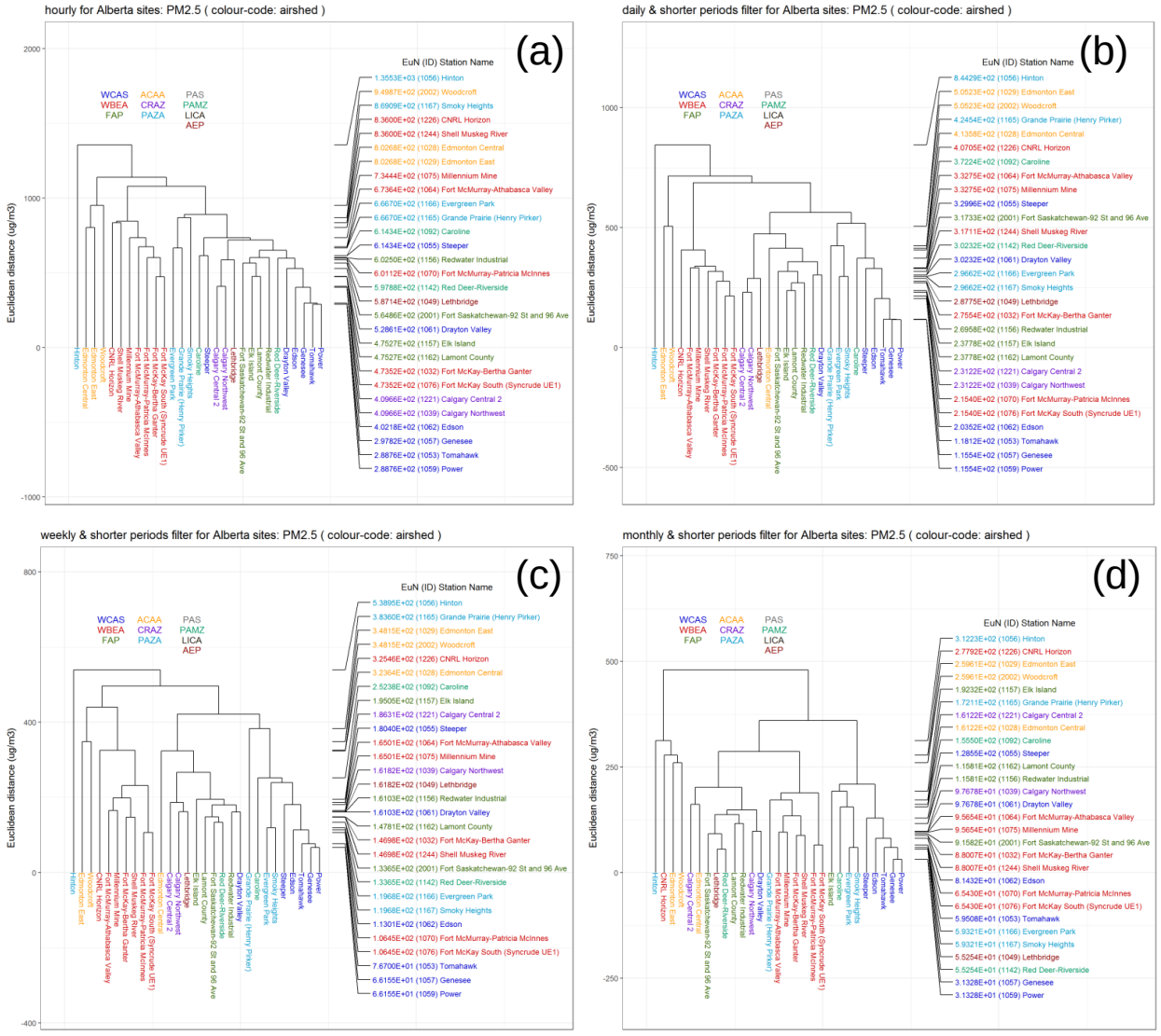


Figure 3.26 Continuous $PM_{2.5}$ Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

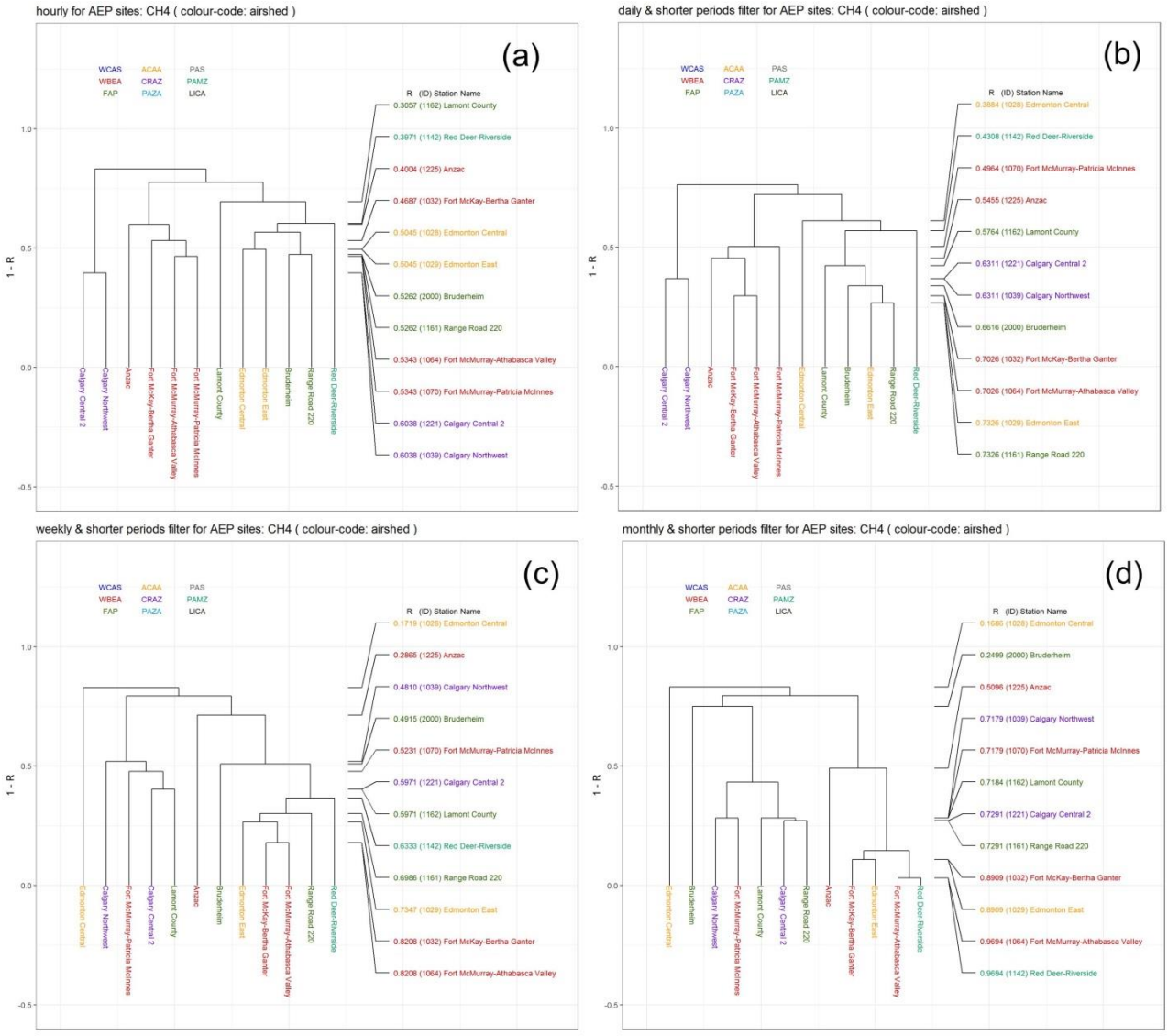


Figure 3.27 Continuous CH_4 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

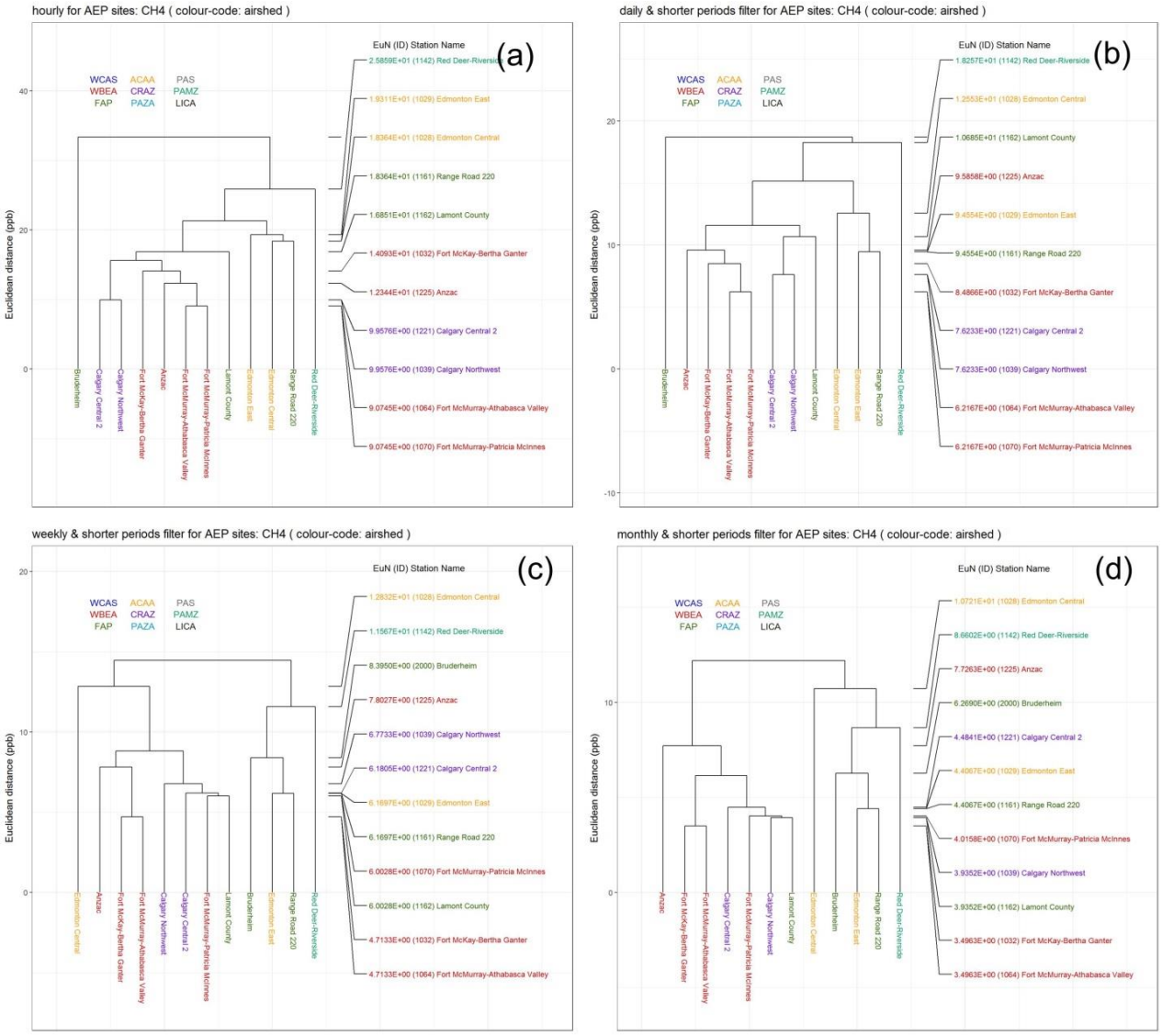


Figure 3.28 Continuous CH_4 Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

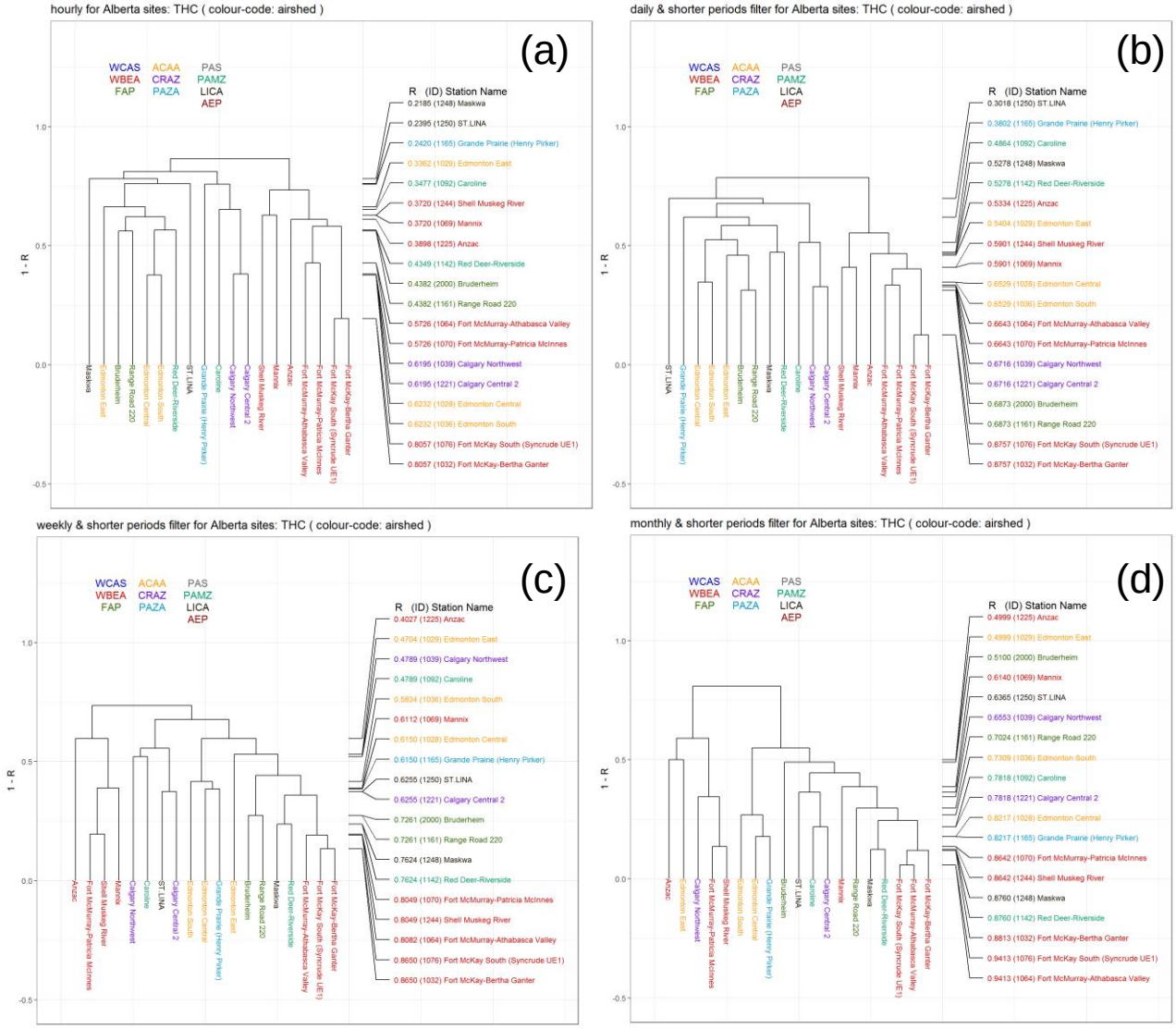


Figure 3.29 Continuous THC 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

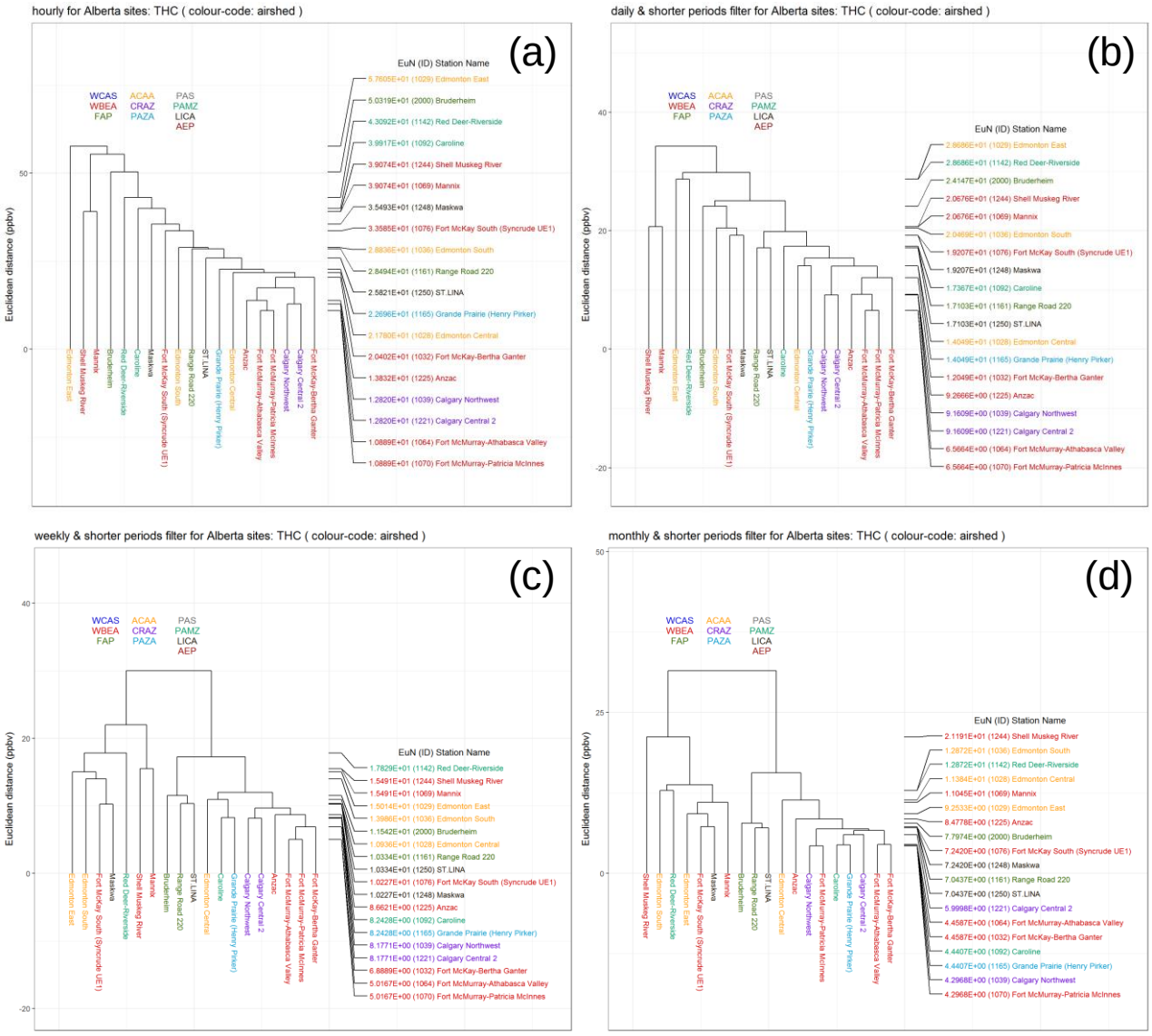


Figure 3.30 Continuous THC Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

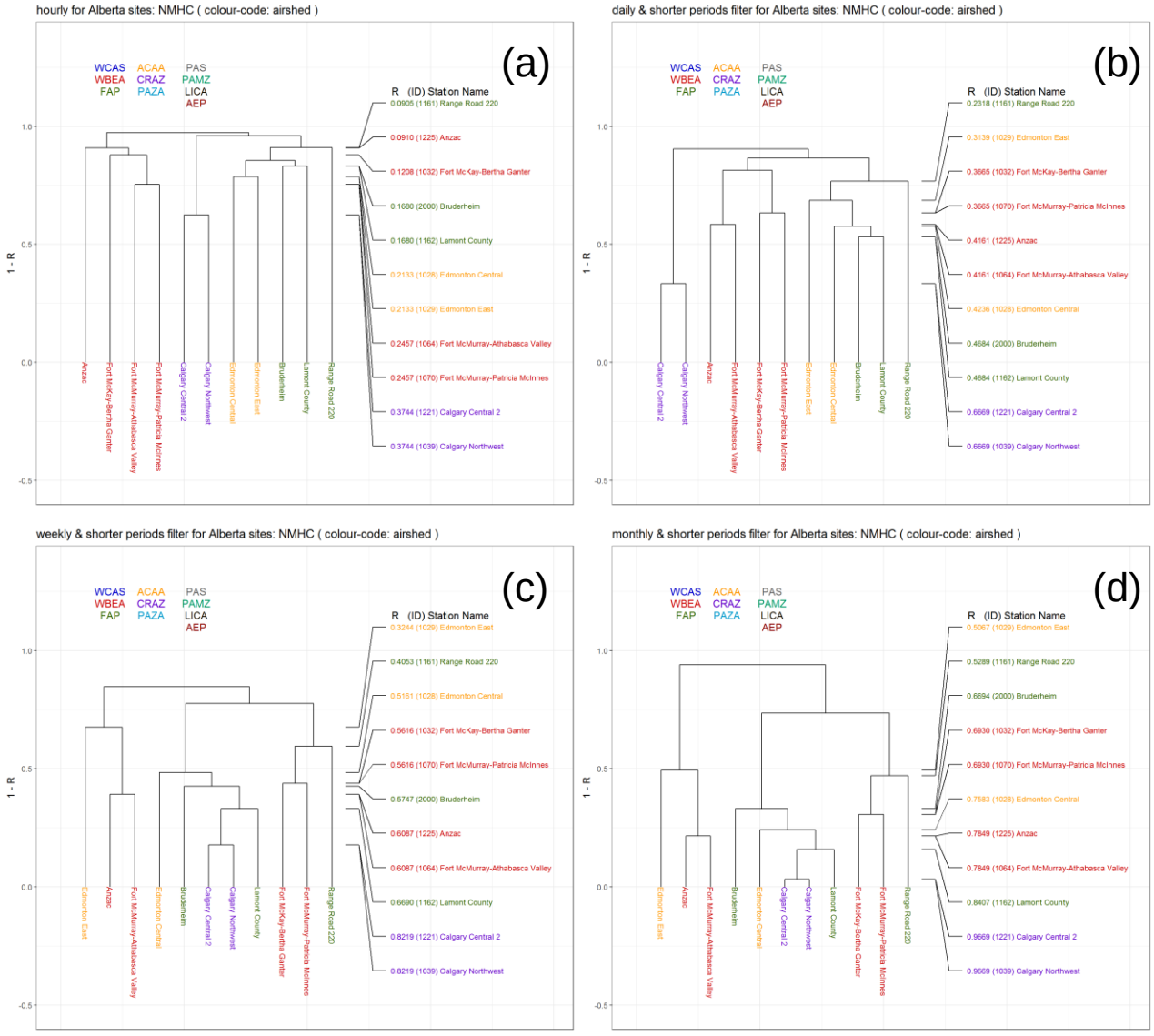


Figure 3.31 Continuous NMHC 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

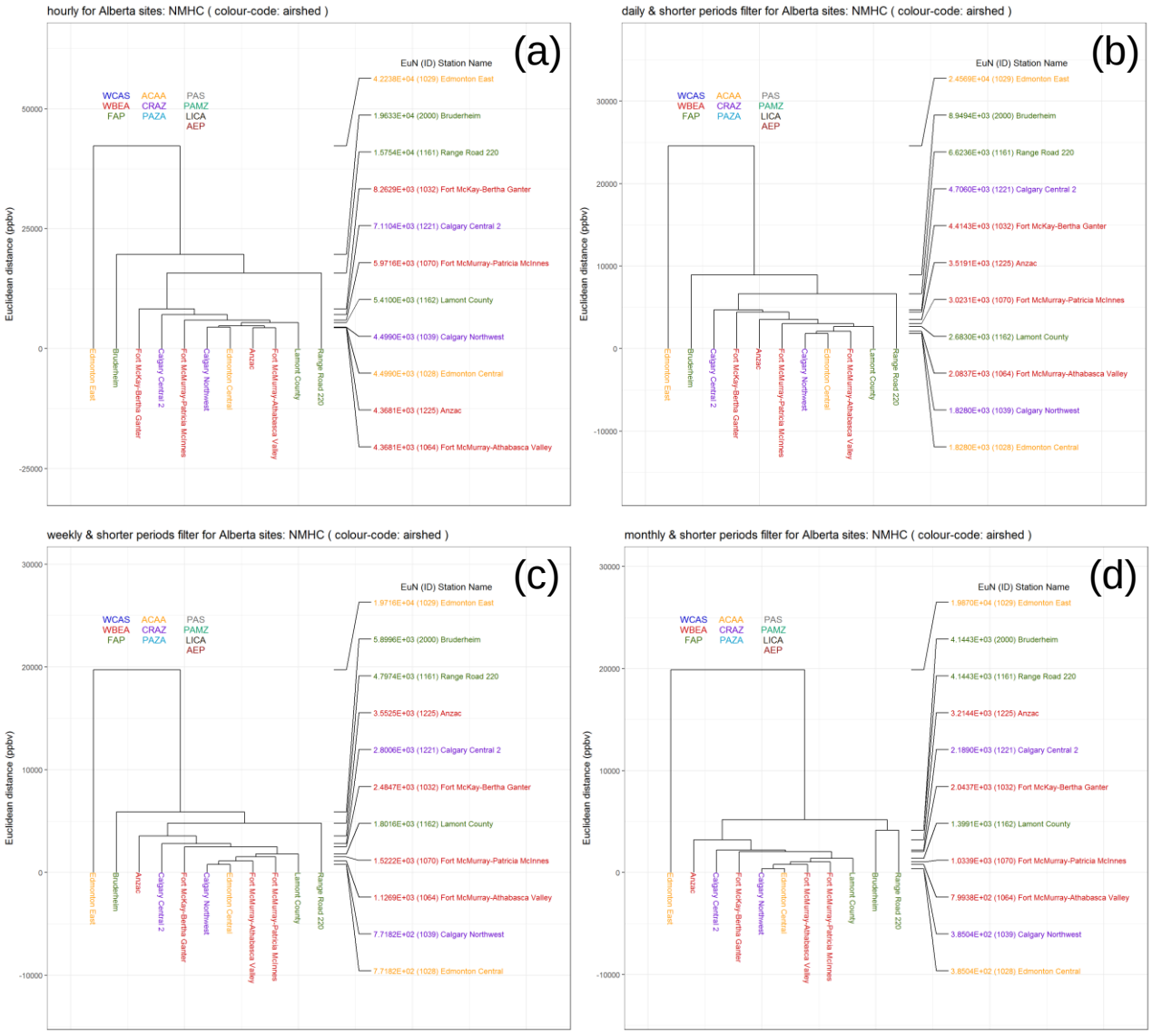


Figure 3.32 Continuous NMHC Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

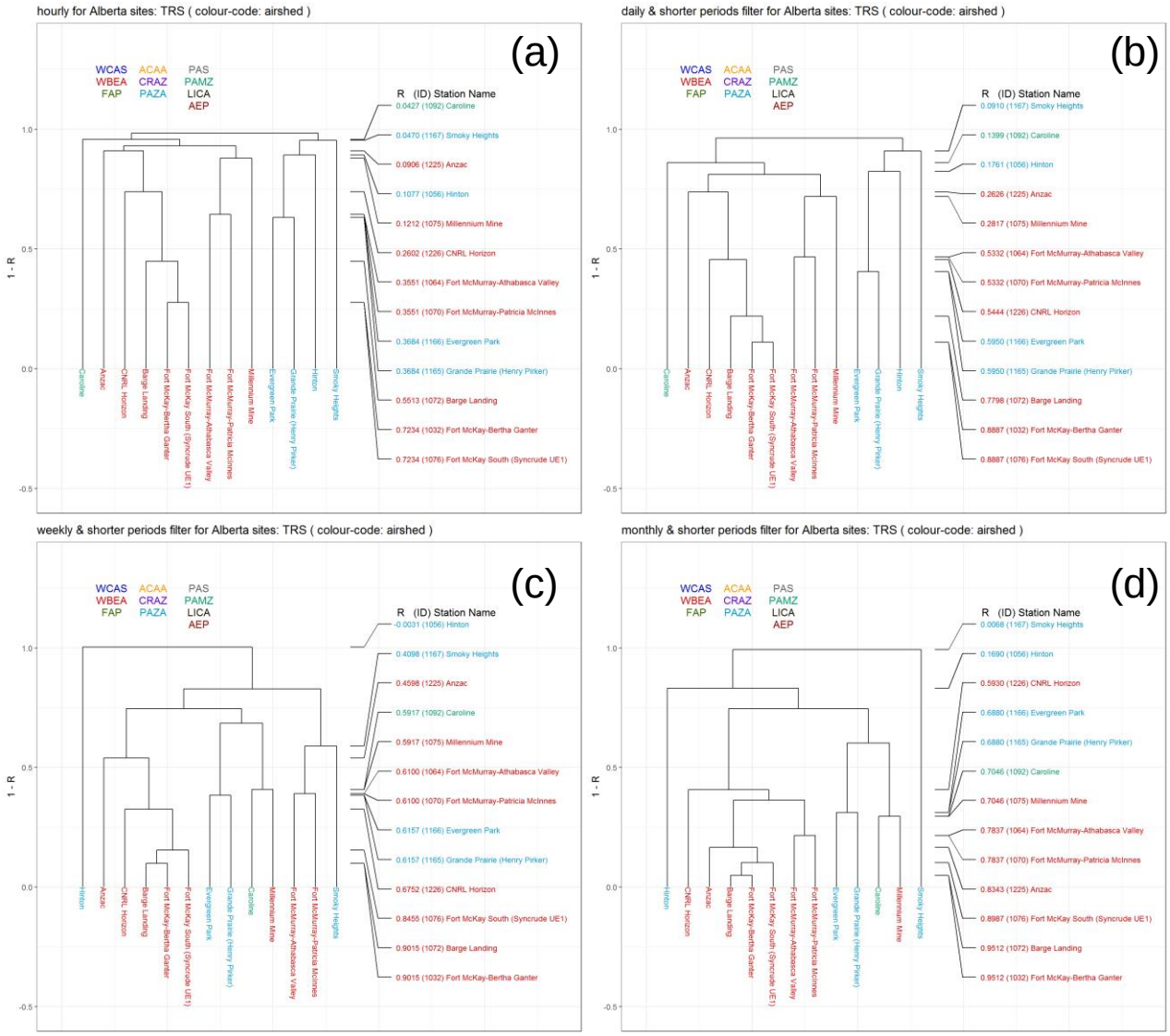


Figure 3.33 Continuous TRS 1-R dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

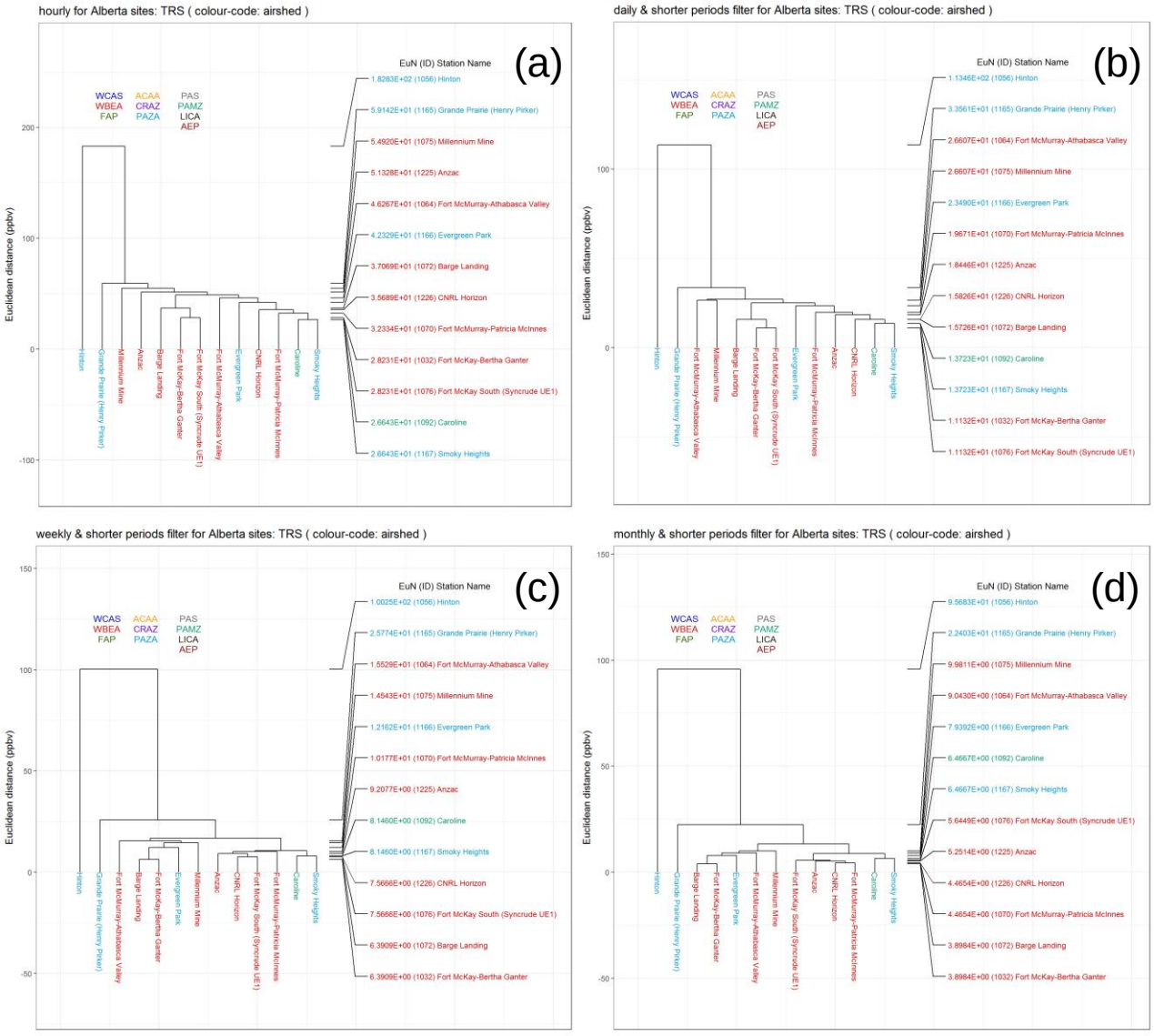


Figure 3.34 Continuous TRS Euclidean distance dendrogram analysis. Panel ordering, Airshed names and colour-coding as in Figure 3.15.

4 Discussion

4.1 Assessing Redundancy

4.1.1 WBEA: Passive and Continuous Monitors

The NO_2 dendrograms depicted in Figure 3.2 clearly separate the continuous monitors from the passives on the basis of correlation. In past uses of hierarchical clustering for air pollution network analysis (Solazzo and Galmarini, 2015), correlation differences such as those displayed in Figure 3.2 were assumed to represent differences in monitoring network methodology. However, the analysis suggests that the reported continuous and passive monitor bimonthly averages are sufficiently different that they do not correlation cluster according to location, but rather cluster according to measurement technology. Collocated sites such as the continuous and passive monitors at Fort McKay-Bertha Ganter (1032C, 1032P) cluster more closely with other continuous or passive monitors than with each other, further demonstrating that the two technologies are not providing equivalent observations. The differences may relate to the precision of the instrumentation and the frequency of low concentration observations – this possibility is examined in Section 4.3, where air quality model output is used as a surrogate to examine the potential issues related to random error in the sampling methodology (precision).

The continuous and passive monitors Euclidean distance dendrogram (Figure 3.3) shows a large degree of spatial variability in concentrations measured, with clustering for three continuous stations splitting at the 24 ppbv level, and the remaining continuous monitors splitting from the rest at the 14 ppbv level. Meanwhile, the corresponding 1-R dendrogram (Figure 3.2) shows a relatively high correlation between Shell Muskeg River/1244C, Fort McMurray Athabasca Valley/1064C and Millennium Mine/1075C. This pattern of high correlations versus high Euclidean distances for the different continuous monitors is explained based on physical proximity between these monitors and the local emissions sources; for example, Fort Chipewyan/1071C differs from the other stations significantly for both metrics, and this would be expected, since that station is far (~ 200 km) from the main emissions region of the Athabasca oil sands. The relatively high correlation between Shell Muskeg River/1244C, Fort McMurray Athabasca Valley/1064C and Millennium Mine/1075C may represent times when the sources at the centre of the emissions region (1075C) are being transported northwards (1244C) or southwards (1064C) along the river valley at low levels (possibly with inversions, given the difference in Euclidean distances between valley-bottom 1064C and adjacent but elevated 1070C). The high values of the Euclidean norm for these three stations would thus represent the specific high concentration transport events from the central emissions region to the periphery. Passive monitors show, in general, smaller magnitudes for the Euclidean distance than the continuous monitors. This results from the passive sampler's inability to resolve short-term concentration peaks (see Section 2.2.1, and references quoted therein).

The above observation and results provided in Section 3.1 suggest that the relationship between the clustering metrics and the pollutant being studied may be complex, and may depend on several factors. Factors could include, for example, whether the pollutant sources are broadly distributed over a large surface area or concentrated in a stack emission, the extent to which emissions have become dispersed

downwind, and the relative rate of uptake of the emitted pollutants or their products by deposition. One may nevertheless use the clustering to rank stations based on their degree of similarity. Stations which join clusters at low levels of *dis*similarity (i.e., high levels of similarity, or high correlation coefficients and low Euclidean distances) are potentially more redundant than stations with higher levels of dissimilarity. Figures 3.2(a) and 3.3(a) include a ranking on the right side of each panel of these figures, and the resulting numbers are also included in Table 4.1. For each metric and species, the data from stations appearing at the *bottom* of the table are the most *similar*, and hence one measure of their level of redundancy, with regards to that specific metric and species examined, and not taking into account other factors as outlined in section 1.2 of this report. We note that the long-term averaging employed here should not be used to determine relative levels of redundancy with regards to continuous stations due to their ability to provide information at other time scales, and the reader is directed to Section 4.1.4 for the companion analysis of continuous data for an analysis at multiple time scales.

Table 4.1 WBEA Bimonthly NO_2 Similarity Ranking. Note that stations at the bottom of the two columns are the most similar (hence one measure of their level of redundancy) with respect to each metric of dissimilarity.

R	station ID	station name	EuN	station ID	station name
0.55	9913P	JP213	21.74	1075C	Millennium Mine
0.68	1071C	Fort Chipewyan	15.45	1244C	Shell Muskeg River
0.74	9912P	JP212	15.45	1064C	Fort McMurray-Athabasca
0.78	9910P	JP205	10.60	9912P	JP212
0.79	9919P	AH7	9.48	9920P	R2
0.82	9903P	BM10	9.12	9908P	JP104
0.86	9914P	NE10	9.12	1032P	Fort McKay-Bertha Ganter
0.86	9921P	SM7	6.74	1226C	CNRL Horizon

0.86	9902P	AH8	6.10	1070C	Fort McMurray-Patricia McInnes
0.86	9907P	JP102	5.76	1032C	Fort McKay-Bertha Ganter
0.87	9916P	NE7	5.76	1076C	Fort McKay South (Syncrude)
0.87	9920P	R2	5.38	9907P	JP102
0.88	9904P	BM11	5.38	1225C	Anzac
0.88	9905P	BM7	4.78	9909P	JP107
0.89	9908P	JP104	4.30	9916P	NE7
0.89	9909P	JP107	4.24	9919P	AH7
0.89	9915P	NE11	3.61	9915P	NE11
0.89	1075C	Millennium Mine	3.57	1071C	Fort Chipewyan
0.91	1032P	Fort McKay-Bertha Ganter	3.25	9901P	AH3
0.92	9906P	JP101	3.09	9902P	AH8
0.93	1064C	Fort McMurray-Athabasca	1.96	9911P	JP210
0.93	1244C	Shell Muskeg River	1.96	9906P	JP101
0.95	1226C	CNRL Horizon	1.88	9910P	JP205

0.96	9901P	AH3	1.41	9921P	SM7
0.96	9911P	JP210	1.21	9913P	JP213
0.97	1225C	Anzac	1.18	9914P	NE10
0.98	1070C	Fort McMurray-Patricia McInnes	1.12	9903P	BM10
0.99	1032C	Fort McKay-Bertha Ganter	1.10	9905P	BM7
0.99	1076C	Fort McKay South (Syncrude)	1.10	9904P	BM11

It can be seen from Table 4.1 that the choice of the most relevant metric for the monitoring network may have a key role in determining which stations may be considered potentially redundant[2] ; the ranking in this particular case differs (and is almost reversed) depending on whether 1-R or Euclidean distance is used to rank stations. Stations which have relatively high correlation (low 1-R) values may have relatively high Euclidean distances, and stations which have relatively low correlations (high 1-R) may have relatively low Euclidean distances.

The metrics measure the similarity of different aspects of the data records. The 1-R metric assesses similarity on the basis of the variation in concentration over time, while the Euclidean distance metric assesses similarity on the basis of differences in concentration magnitude. If the former quality is more important with regards to the intended purpose of the monitoring, then the left (1-R) column of Table 4.1 takes precedence, and stations with the higher R values (bottom of the left column) would be considered more redundant. If on the other hand, the stations with the most similar concentrations being reported are considered the most redundant, then the right-hand column of Table 4.1 takes precedence, and stations with the smaller Euclidean distances should be considered to be the most redundant. We note that only data from two stations in Table 4.1 *both* fall within list of ten lowest 1-R and Euclidean metric values (that is, might be considered redundant for both 1-R and Euclidean distances); NO_2 bimonthly concentrations from passive stations JP101/9906P and JP210/9911P.

[2] We note again that the passive and continuous monitoring records were binned to the same bimonthly interval to determine the level of comparability between the two different types of instrumentation and methodology. The relative similarity levels and redundancies for Table 4.1 are with respect to the two month averaging time. The continuous monitors provide information down to hourly time-scales. Table 2.1 only assesses their potential relative redundancy for two month averaging - the analyses appearing in sections 3.4 and 4.1.4 should have precedence in assessing relative redundancy for continuous monitors.

This dichotomy between the two metrics is illustrated further in Table 4.2, where the clusters with the highest correlation coefficients are compared to the Euclidean distances between the members of 1-R clusters. Figure 3.2 shows that at correlation level R = 0.9 (1-R = 0.1), three clusters of more than one station are formed, two of which cluster continuous stations and one cluster of passive monitors. From left to right: the first R = 0.9 cluster consists of Shell Muskeg River/1244C and Fort McMurray-Athabasca Valley/1064C continuous monitors; the second of Anzac/1225C, Fort McMurray-Patricia McInnes/1070C, Fort McKay South/1076C, Fort McKay-Bertha Ganter/1032C, and CNRL Horizon/1226C continuous monitors; and the third of Fort McKay-Bertha Ganter/1032P, JP101/9906P, AH3/9901P and JP210/9911P passive monitors. The Euclidean distance between each of the members of these R=0.9 clusters are given in Table 4.2 (the values in the table were obtained via tracing the Euclidean distance dendrogram to find the Euclidean distance levels where these stations connect). The physical distances between the stations are given in Table 4.3 for reference.

Table 4.2 Euclidean distance (ppb) for NO_2 clusters at dissimilarity level 1-R = 0.1: Fort McMurray-Athabasca Valley (1064C), Shell Muskeg River (1244C), Fort McKay-Bertha Ganter (1032C), Fort McMurray-Patricia McInnes (1070C), Fort McKay South (1076C), Anzac (1225C), and CNRL Horizon (1226C), Forth McKay-Bertha Ganter (1032P), JP101 (9906P), AH3 (9901P) and JP210 (9911P).

cluster	station ID	1064C	1244C			
1	1064C 1244C	0 15.5	15.5 0			
cluster	**station ID**	**1032C**	**1070C**	**1076C**	**1225C**	**1226C**
2	1032C 1070C 1076C 1225C 1226C	0 6.1 5.8 6.7 14.4	14.4 0 6.1 14.4 6.7	14.4 6.1 0 14.4 6.7	6.7 14.4 14.4 0 14.4	14.4 6.7 6.7 14.4 0
cluster	**station ID**	**1032P**	**9901P**	**9906P**	**9911P**	**9911P**
3	1032P 9901P 9906P 9911P	0 10.7 10.7 10.7	10.7 0 3.2 3.2	10.7 0 3.2 3,2	10.7 3.2 0 2.0	10.7 3.2 2.0 0

Table 4.3 Distance (km) for NO_2 clusters at dissimilarity level 1-R = 0.1: Fort McMurray –Athabasca Valley (1064C), Shell Muskeg River (1244C), Forth McKay-Bertha Ganter (1032C), Fort McMurray-Patricia McInnes (1070C), Forth McKay South (1076C), Anzac (1225C), and CNRL Horizon (1226C), Forth McKay-Bertha Ganter (1032P), JP101 (9906P), AH3 (9901P) and JP210 (9911P).

cluster	station ID	1064C	1244C			
1	1064C	0	58			
	1244C	58	0			
cluster	**station ID**	**1032C**	**1070C**	**1076C**	**1225C**	**1226C**
2	1032C	0	50	4	90	14
	1070C	50	0	45	43	64
	1076C	4	45	0	86	18
	1225C	90	43	86	0	104
	1226C	14	64	18	104	0
cluster	**station ID**	**1032P**	**9901P**	**9906P**	**9911P**	
3	1032P	0	63	82	125	
	9901P	63	0	73	63	
	9906P	82	73	0	117	
	9911P	125	63	117	0	

The members of the first highly correlated cluster listed in Table 4.2 have Euclidean distances as high as 15.5 ppbv; despite the monitors having a similar shape of their time series, the difference in their magnitudes is considerable.

For the second cluster, the Euclidean distance dendrogram shows that Anzac/1225C is clearly different from the other stations, with a value of 14.2 ppb, while the remaining stations have lower values of between 5.8 and 6.7 ppb. Fort McKay South/1076C and Fort McKay-Bertha Ganter/1032C are separated by ~4 km, and are highly correlated, but their Euclidean norms are separated by 14.4 ppbv in Figure 3.3, indicating a large difference in the magnitudes between these two stations. Fort McMurray-Patricia McInnes/1070C is located at distances over 40km from the other stations in cluster 2, has a high correlation level, but has Euclidean distances of 14.4 ppbv with 1032C to the north, and 6.1 to 6.7 with the stations to the south.

For the third cluster, the Euclidean distance dendrogram shows the Forth McKay-Bertha Ganter/1032P station departs from the other stations at a level of 10.7 ppb; the remaining stations are closer to each other in terms of Euclidean distance but none of these stations are physically close to each other (see Figure 3.1, Figure 3.6 and Table 4.3). This suggests a lack of precision in the passive monitors, particularly given the lack of correlation between collocated continuous and passive monitors 1032C and 1032P; 1032P correlates more highly with other passive monitors that are between 63 and 125 km

distant, than with the collocated continuous monitor 1032C. The Euclidean distance between collocated 1032C and 1032P is 14.4 ppb, while the Euclidean distances between 1032P and the more distant stations making up cluster 3 in Table 4.2 are smaller; 10.7, 3.2 and 2.0 ppb. For both 1-R and Euclidean distance metrics, station 1032P is more “like” distant passive stations than an adjacent continuous station.

Monitors that are located at the same site such as Fort McKay-Bertha Ganter passive and continuous monitors (1032C, P) would be expected to give similar results but they are not correlating significantly, nor show smaller levels of Euclidean distance. In contrast with this situation, Fort McKay South/1076C, at a distance of about 4 km from the latter stations, shows high level of correlation to 1032C, as might be expected. Publication such as Bari et al (2015) and EPCM (2000) report that Fort McKay-Bertha Ganter and Fort McMurray (passive not included in this analysis due to data limitation) passive monitors report errors up to 15%, with a tendency to underestimate concentration, when compared to continuous monitors.

The dendrograms for SO_2 do not distinguish between continuous and passive monitors to the same extent as for NO_2. However, the resulting SO_2 clusters do not always follow spatial location groupings. There may be some loss of information in the observations associated with the averaging time and the precision of the original observations; this will be discussed later in this chapter of the report. In previous examinations of SO_2 in this region, Bari et al (2015) and EPCM (2000) reported that Fort McKay and Fort McMurray passive monitors (data from the Fort McMurray station was not included in the current analysis due to limited available data) report errors up to 34%, with a tendency to overestimate concentration, when compared to continuous monitors. Here, despite similar magnitudes (Euclidean distance metric), the collocated passive and continuous monitors at Fort McKay-Bertha Ganter are clearly not correlating at high level (1-R=0.54)

Figure 3.4(a) and Figure 3.5 (a) include a ranking on the right side of each panel of these figures, and the resulting numbers are also included in. The dichotomy of rankings between the two different metrics is less than noted for NO_2; some SO_2 monitors fall near the bottom of Table 4.4 for both 1-R and Euclidean distance metrics, indicating a greater degree of potential redundancy for SO_2 for both metrics than for NO_2.

Table 4.4 WBEA Bimonthly SO_2 Similarity Ranking. Note that stations at the bottom of the two columns are the most similar (hence one measure of their level of redundancy with respect to each metric of dissimilarity).

R	station ID	station name	EuN	station ID	station name
0.35	9918P	WF4	6.92	1066C	Mildred Lake
0.44	1066C	Mildred lake	6.31	9918P	WF4
0.54	1032P	Fort McKay-Bertha Ganter	5.97	1069C	Mannix
0.58	9912P	JP212	4.93	9908P	JP104
0.58	9908P	JP104	4.18	9907P	JP102
0.58	9902P	AH8	4.18	1074C	Lower Camp
0.62	9904P	BM11	3.87	9906P	JP101
0.62	9903P	BM10	3.77	9912P	JP212
0.62	1226C	CNRL Horizon	3.42	1244C	Shell Muskeg River
0.67	1075C	Millennium Mine	3.33	9917P	SM8
0.67	1068C	Buffalo Viewpoint	3.24	1032P	Fort McKay-Bertha Ganter
0.67	1074C	Lower Camp	3.13	9902P	AH8
0.67	1064C	Fort McMurray-Athabasca Valley	3.08	1226C	CNRL Horizon
0.68	9915P	NE11	2.98	1075C	Millennium Mine
0.68	1244C	Shell Muskeg River	2.98	1068C	Buffalo Viewpoint

0.73	9909P	JP107	2.95	9909P	JP107
0.77	9907P	JP102	2.89	1064C	Fort McMurray-Athabasca Valley
0.77	1069C	Mannix	2.79	9904P	BM11
0.80	9901P	AH3	2.66	9916P	NE7
0.82	9913P	JP213	2.66	9915P	NE11
0.83	1032C	Fort McKay-Bertha Ganter	2.56	9910P	JP205
0.83	1076C	Fort McKay South (Syncrude UE1)	2.37	1032C	Fort McKay-Bertha Ganter
0.85	1070C	Fort McMurray-Patricia McInnes	2.37	1076C	Fort McKay South (Syncrude UE1)
0.85	1225C	Anzac	2.17	9901P	AH3
0.85	9911P	JP210	2.17	1070C	Fort McMurray-Patricia McInnes
0.85	9906P	JP101	2.12	9913P	JP213
0.86	9914P	NE10	1.87	9903P	BM10
0.87	9916P	NE7	1.73	9911P	JP210
0.87	9910P	JP205	1.73	1225C	Anzac
0.89	9905P	BM7	1.45	1071C	Fort Chipewyan
0.93	9917P	SM8	1.28	9914P	NE10
0.93	1071C	Fort Chipewyan	1.28	9905P	BM7

4.1.2 LICA Passive and Continuous Monitors

The summary table of 1-R and Euclidean distance rankings of the LICA stations for NO_2 are shown in Table 4.5. Similar to the WBEA table (Table 4.1), the station rankings differ between the two metrics, though seven passive stations appear in both columns for the ten “most similar” stations: Fishing Lake/1191P, Dupre/1182P, Lake Eliza/1178P, Muriel-Kehiwin/1181P, Therien/1176P, Fort George/1195P, and La Corey/1183P. Comparing the Euclidean distances for LICA and WBEA stations however shows that the LICA stations tend to have larger Euclidean distances (minimum value LICA: 1.76 vs WBEA 1.10); there is more variation in concentrations between the LICA stations compared to between WBEA stations. In both cases the rankings are relative to the other stations within the given Airshed – here we do not set a specific level for data similarity, but note that the relative rankings are Airshed-specific.

The Cold Lake oil sands area has three stations continuously monitoring both NO_2 and SO_2. The NO_2 bimonthly dendrogram (Figure 3.7) shows that the continuous monitors cluster at a high correlation/low1-R level of dissimilarity (R~ 0.9, 1-R=0.1) and the passives cluster together, with exception of Primrose/1186P and St. Lina/1252P, at a correlation level 0.79. As noted before, the collocated continuous and passive samplers at St. Lina (1250C, 1252P) and Cold Lake South (1174C, 1193P, 1227P) and Maskwa (1248C, 1187P) do not cluster with each other between the two measurement technologies, indicating that the two methodologies are not providing comparable bimonthly average concentrations. Two of LICA’s passive NO2 monitors behave differently from the rest of the monitors (Primrose/1186P and St. Lina/1252P), with 1-R node separation from the rest of the stations at a correlation level of 0.03 and 0.56, respectively. As noted earlier, the time series of Primrose has a single isolated high value not recorded at the other stations, and St. Lina is located upwind of the other stations in the airshed, possibly explaining the lower levels of clustering for these stations. The metric values of LICA’s several collocated passive and continuous monitors vary in magnitude: e.g. St Lina’s link at 1-R=0.56 and Euclidean distance 3.9 ppb; Maskwa’s at 1-R=0.78 and 6.5 ppb; and Cold Lake South’s at 1-R=0.78 and 10.1 ppb.

Table 4.5 LICA Bimonthly NO_2 Similarity Ranking. Note that stations at the bottom of the two columns are the most similar (hence one measure of their level of redundancy) with respect to each metric of dissimilarity.

R	station ID	station name	EuN	station ID	station name
0.03	1186P	Primrose	15.59	1186P	Primrose
0.56	1252P	St. Lina	10.50	1199P	Town of Bonnyville
0.82	1192P	Beaverdam	10.15	1174C	Cold Lake South
0.84	1199P	Town of Bonnyville	5.66	1252P	St. Lina
0.88	9919P	Frog Lake	5.31	1248C	Maskwa
0.90	1174C	Cold Lake South	3.92	1250C	ST.LINA
0.91	1177P	Flat Lake	3.90	1190P	Clear Range
0.92	1190P	Clear Range	3.24	1227P	Cold Lake South Passive 2
0.93	1187P	Maskwa	3.24	1193P	Cold Lake South Passive
0.93	1248C	Maskwa	3.07	1189P	Frog Lake
0.93	1250C	ST.LINA	2.53	1195P	Fort George
0.93	1191P	Fishing Lake	2.53	1183P	La Corey
0.94	1182P	Dupre	2.43	1187P	Maskwa
0.95	1178P	Lake Eliza	2.41	1192P	Beaverdam
0.96	1193P	Cold Lake South Passive	2.30	1182P	Dupre
0.96	1227P	Cold Lake South Passive 2	2.30	1176P	Therien
0.96	1176P	Therien	1.90	1177P	Flat Lake
0.96	1181P	Muriel-Kehiwin	1.87	1191P	Fishing Lake
0.96	1183P	La Corey	1.76	1181P	Muriel-Kehiwin
0.96	1195P	Fort George	1.76	1178P	Lake Eliza

The differences between 1-R and Euclidean distances are explored for the most highly correlated stations in Table 4.6 (1-R < 0.1; R> 0.9) matched with their Euclidean distance between the different stations belonging to a specific 1-R <0.1 cluster. Table 4.7 shows the corresponding physical distances between the stations.

Table 4.6 Euclidian distance (ppb) for NO2 clusters at dissimilarity level 1-R = 0.1: Maskwa (1248C), St. Lina (1250C), Le Corey (1183P), Fishing Lake (1191P), Fort George (1195P), Therien (1176P), Lake Eliza (1178P), Maskwa (1187P), Cold Lake South Passive 1 and 2 (1193P, 1227P), Flat Lake (1177P), Muriel-Kehiwin (1181P), Dupre (1182P), and Clear Range (1190P).

cluster	station ID	1248C	1250C			
1	1048C	0	6.5			
	1250C	6.5	0			
cluster	**station ID**	**1183P**	**1191P**	**1195P**		
2	1183P	0	6.5	2.5		
	1191P	6.5	0	6.5		
	1195P	2.5	6.5	0		
cluster	**station ID**	**1178P**	**1187P**	**1193P**	**1227P**	
3	1178P	0	2.7	3.8	3.8	
	1187P	2.7	0	3.8	3.8	
	1193P	3.8	3.8	0	3.2	
	1227P	3.8	3.8	3.2	0	
cluster	**station ID**	**1176P**	**1177P**	**1181P**	**1182P**	**1190P**
4	1176P	0	2.7	2.7	2.3	3.9
	1177P	2.7	9	1.9	2.7	3.9
	1181P	2.7	1.9	0	2.7	3.9
	1182P	2.3	2.7	2.7	0	3.9
	1190P	3.9	3.9	3.9	3.9	0

Table 4.7 Distance between stations (km) for NO_2 clusters at dissimilarity level 1-R = 0.1: Maskwa (1248C), St. Lina (1250C), Le Corey (1183P), Fishing Lake (1191P), Fort George (1195P), Therien (1176P), Lake Eliza (1178P), Maskwa (1187P), Cold Lake South Passive 1 and 2 (1193P, 1227P), Flat Lake (1177P), Muriel-Kehiwin (1181P), Dupre (1182P), and Clear Range (1190P).

cluster	station ID	1248C	1250C			
1	1048C	0	81			
	1250C	81	0			
cluster	**station ID**	**1183P**	**1191P**	**1195P**		
2	1183P	0	82	69		
	1191P	82	0	44		
	1195P	69	44	0		
cluster	**station ID**	**1178P**	**1187P**	**1193P**	**1227P**	
3	1178P	0	99	90	90	
	1187P	99	0	26	26	
	1193P	90	26	0	0	
	1227P	90	26	0	0	
cluster	**station ID**	**1176P**	**1177P**	**1181P**	**1182P**	**1190P**
4	1176P	0	27	40	29	110
	1177P	27	0	30	40	90
	1181P	40	30	0	27	71
	1182P	29	40	27	0	96
	1190P	110	90	71	96	0

The cluster with continuous monitors shows the highest Euclidean and physical distance values between the stations. Bimonthly average NO_2 concentrations from those continuous monitoring stations with high correlation levels (1-R<0.1) had relatively high Euclidean distances (different magnitudes), possibly indicating differences in downwind distance from similar sources, or other factors, such as large scale (both temporally and spatially) variation in the meteorological conditions, similar operating cycles for the different facilities monitored, etc.

The summary of 1-R and Euclidean distance rankings for bimonthly SO_2 are shown Table 4.8. Eight SO_2 stations Cold Lake South Passive/1193P, Cold Lake South Passive2/1227P, La Corey/1183P, Dupre/1182P, Fort George/1195P, Clear Range/1190P, Lake Eliza/1178P, and Flat Lake/1177P, are all within the ten lowest 1-R and Euclidean distances. Table 4.9 and Table 4.10 show the corresponding Euclidean distances, and the physical distance between the stations, for the clusters at dissimilarity level 0.1. Unlike NO_2, the magnitudes of the Euclidean distances (Table 4.9) for the lower scoring LICA

stations for this metric are lower than their WBEA counterparts; the LICA stations are more similar to each other in terms of Euclidean distance than for WBEA.

Table 4.8 LICA Bimonthly SO_2 Similarity Ranking. Note that stations at the bottom of the two columns are the most similar (hence one measure of their level of redundancy) with respect to each metric of dissimilarity.

R	station ID	station name	EuN	station ID	station name
0.02	1199P	Town of Bonnyville	3.89	1250C	St. Lina
0.32	1174C	Cold Lake South	3.39	1174C	Cold Lake South
0.54	1179P	Telegraph Creek	2.22	1248C	Maskwa
0.54	1198P	Hilda Lake	2.16	1187P	Maskwa
0.60	1250C	St. Lina	2.16	1198C	Hilda Lake
0.60	1248C	Maskwa	2.05	1199P	Town of Bonnyville
0.81	1186P	Primrose	1.57	1186P	Primrose
0.81	1187P	Maskwa	1.56	1252P	St. Lina
0.85	1176P	Therien	1.56	1197P	Mahihkan
0.85	1252P	St. Lina	1.22	1179P	Telegraph Creek
0.86	1192P	Beaverdam	1.02	1189P	Frog Lake
0.86	1191P	Frog Lake	0.87	1176P	Therien
0.88	1197P	Mahihkan	0.80	1177P	Flat Lake
0.88	1183P	La Corey	0.79	1192P	Beaverdam
0.88	1182P	Dupre	0.73	1178P	Lake Eliza
0.89	1178P	Lake Eliza	0.73	1190P	Clear Range
0.91	1193P	Cold Lake South Passive	0.61	1183P	La Corey
0.91	1195P	Fort George	0.50	1227P	Cold Lake South Passive 2
0.91	1177P	Flat Lake	0.50	1193P	Cold Lake South Passive

Some of the LICA SO_2 monitors show a substantially different behavior (high values of 1-R; low R values) compared to the other LICA stations (Figure 3.9): the passive Town of Bonnyville/1199P, and all three continuous monitors, have larger relative dissimilarity with respect to the remaining passive monitors, with correlation levels at or below R=0.6. There is a moderately dissimilar cluster of passive monitors located at Hilda Lake/1198P, Telegraph Creek/1179P linking at correlation level 0.54. Passive and continuous monitors located at the same sites link only at high levels of 1-R dissimilarity and have some of the highest Euclidean distances of the LICA SO_2 analysis. The sites tend to correlate at levels lower than 0.5 and for Euclidean distances, 1.5 ppbv for St Lina (1250C, 1252P), 5.2 ppbv for Maskwa (1248C/1187P), and 3.4 ppbv for Cold Lake South's monitors (1174C, 1193P, 1227P).

Table 4.9 and Table 4.10 show the Euclidean distance and the physical distance between stations clustering with a 1-R value of 0.10 (R=0.90). The Euclidean distances are often below 1 ppb, though the spatial separation between the stations is sometimes large between these passive stations.

Table 4.9 Euclidean distance (ppb) for SO_2 clusters at dissimilarity level 1-R = 0.1: Fort George (1195P) and Flat Lake (1177P), Cold Lake South Passive 1 and 2 (1193P, 1227P).

cluster	station ID	1195P
1	1177P	1.5
cluster	**station ID**	**1227P**
2	1193PC	0.5

Table 4.10 Distance between stations (km) for SO_2 clusters at dissimilarity level 1-R = 0.1: Dupre (1182P), Fort George (1195P), Flat Lake (1177P), Lake Eliza (1178P), Beaverdam and Frog Lake, Cold Lake South Passive 1 and 2 (1193P, 1227P).

cluster	station ID	1195P
1	1177P	43
cluster	**station ID**	**1227P**
2	1193PC	0

4.1.3 All Alberta NO2 and SO2 Passive and Continuous Monitors

Evaluation of redundancies on a provincial basis for the bimonthly passive and continuous monitors was carried out in order to assist in potential decision making across Airsheds on passive monitoring site redundancies. For example, stations managed by different Airsheds may be in sufficiently close proximity that they may be observing similar sources, and if sufficiently similar and in close proximity, could be considered redundant. Decisions on redundancies must be made carefully however, for several reasons.

First, continuous stations were included with passive stations in order to be able to determine the level of comparability (similarity) of the different methodologies for observations – but the higher time resolution of the continuous monitors allows them to be used for additional purposes besides long-term averaging and hence their level of similarity for bimonthly averages will be less relevant in determining the relative level of redundancy for these stations. Next, as noted earlier, station time series with very similar correlation coefficients may represent the impact of sources with a similar temporal emissions pattern across the province – the rush hour peaks of NO_2 in the morning and afternoon, for urban to suburban regions where the dominant source of NO_2 will be mobile emissions, will result in high levels of 1-R similarities between urban stations in different cities, despite being influenced by different local emission sources. A similar effect could be expected to occur for SO_2 stations influenced by widely spatially separated coal-fired powerplants with the same daily and seasonal cycle of power output. Physical proximity between stations should thus be considered in assessing redundancies based on either the 1-R or Euclidean distance metrics. Station time series may be highly similar due to their close proximity for either metric (and hence greater likelihood of being influenced by the same sources at the same time), due to being far apart yet influenced by separate emissions sources which happen to have a similar temporal variation (1-R metric), or being located far apart and have sufficiently low concentrations that they have relatively high similarities (Euclidean distance metric) since they are sampling background air. For the Euclidean distance metric, physical proximity should also be considered – stations may be highly similar with this norm due to being (a) located close together and measuring concentrations of pollutants associated with the same sources, (b) located far apart, and measuring concentrations of pollutants associated with different sources which happen to have the same source strength, (c) located far apart, and near to no major sources, so that the Euclidean distances are uniformly low since the stations are all observing low concentration, “background”, air.

Passives have been used as an alternative to continuous monitors for monitoring temporal trends of air pollutants in remote areas (Krupa and Legge, 2000; Cox, 2003; Seethapathy et al., 2008; Bytnerowicz et al., 2010) and evaluating air quality of large areas (Gerboles et al., 2006). However, passive sampling disadvantages compared to continuous are low sensitivity, inability to resolve concentrations peaks, and adverse effects of meteorological conditions (Tang et al. 1997, 1999; Krupa and Legge, 2000; Tang, 2001; Kirby et al., 2001; Partyka et al., 2007; Fraczek et al., 2009; Salem et al., 2009; Zabiegala et al., 2010). Moreover, monthly meteorological information needed to calculate the diffusion rate is obtained from the nearest site with meteorological observations, as most passive sampling sites do not have meteorological information. These constraining factors could influence the sampling and, therefore, the accuracy of the results, causing under- or overestimation of ambient gas concentrations in relation to continuous analyzers (Krupa and Legge, 2000). There have been several studies comparing passive and continuous analyzers in Alberta (WBK, 2007; Hsu et al., 2010; Pippus, 2012; Bari et al., 2015). Bari et al. (2015), the study with the highest number of samples compared, cautioned that direct comparisons between NO_2 and SO_2 continuous and passive samplers may be hampered by lower field accuracy in the latter. Several studies show that passive samplers overestimate SO_2 ambient concentrations and underestimate NO_2, in regard to continuous monitors. We note that these comparisons were done for urban sites only; the work which follows includes an objective comparison of passive and continuous monitors for rural, urban, and industrial sites outside of urban regions.

All issues described above must be considered when making use of the 1-R and Euclidean distance metrics rankings for the 126 bimonthly NO_2 passive and continuous stations used in the following analysis, appearing below in Table 4.11. An important feature which may be seen from Table 4.11 is that the stations with the highest correlation coefficients (lowest values of the 1-R dissimilarity metric) are usually not the same stations as the ones with the lowest values of the Euclidean distance (see bottom of Table 4.11). The stations with the highest correlation coefficients are a mixture of passive and continuous monitors, while the stations with the lowest Euclidean distances tend to be passive monitors. For the latter, several are in higher elevation locations and/or appear to be sampling relatively low concentrations (close to zero), e.g. Limestone Mountain/9943P, Parker Ridge/9939P, Bow Summit/9938P. The lowest (< 3 ppb) Euclidean distances may thus reflect stations which are sampling "background" air – the similarity with regards to this metric may be due to the stations having uniformly low concentrations, despite being located in different parts of the province.

Table 4.11 Bimonthly NO_2 Similarity Ranking. Note that stations at the *bottom* of the two columns are the most similar (hence one measure of their level of redundancy) with respect to each metric of dissimilarity.

R	station ID	station name	EuN	station ID	station name
-0.04	1186P	Primrose	27.7	1028C	Edmonton Central
0.12	9942P	Baseline Mountain	20.7	1075C	Millennium Mine
0.35	9939P	Parker Ridge	18.3	1156C	Redwater Industrial
0.38	9938P	Bow Summit	17.5	1029C	Edmonton East
0.43	9933P	PAS 15	15.5	1244C	Shell Muskeg River
0.46	9932P	PAS 14	15.5	1064C	Fort McMurray-Athabasca Valley
0.47	9913P	JP213	15.4	1186P	Primrose
0.55	1172P	PAS 19	13.2	1142P	Red Deer Riverside
0.56	9940P	Bighorn	11.8	1172P	PAS 19

0.57	9934P	PAS 16	11.4	2000C	Bruderheim
0.65	1071C	Fort Chipewyan	11.2	1142C	Red Deer-Riverside
0.67	1049C	Lethbridge	10.7	2001C	Fort Saskatchewan-92St & 96Ave
0.69	9970P	Panther River	10.7	1159C	Ross Creek
0.71	1156C	Redwater Industrial	9.64	1049C	Lethbridge
0.72	1252P	St. Lina	9.47	9920P	R2
0.73	9943P	Limestone Mountain	9.12	1032P	Fort McKay-Bertha Ganter
0.74	9931P	PAS 13	8.83	9936P	PAS 18
0.74	9912P	JP212	8.83	9929P	PAS 11
0.74	1052C	Violet Grove	8.74	1241C	Wagner2
0.76	9922P	PAS 1	8.71	1161C	Range Road 220
0.76	9929P	PAS 11	8.43	1172C	Crescent Heights
0.76	9941P	Sunchild	8.43	1058C	Meadows
0.76	9925P	PAS 4	8.15	1199P	Town of Bonnyville
0.76	9928P	PAS 8	8.07	1039C	Calgary Northwest
0.77	9966P	Ferrybank	8.07	1036C	Edmonton South

0.77	2000C	Bruderheim	7.91	1052C	Violet Grove
0.79	9927P	PAS 7	7.91	9935P	PAS 17
0.79	9930P	PAS 12	7.83	9912P	JP212
0.79	9954P	Bottrel	7.77	1054C	Carrot Creek
0.80	9919P	AH7	7.44	1059C	Power
0.80	9910P	JP205	7.38	9908P	JP104
0.81	1055C	Steeper	7.31	1162C	Lamont County
0.81	1059C	Power	7.25	1053C	Tomahawk
0.81	1241C	Wagner2	6.74	1226C	CNRL Horizon
0.82	9924P	PAS 3	6.46	1174C	Cold Lake South
0.82	9935P	PAS 17	6.36	9926P	PAS 6
0.83	1192P	Beaverdam	6.35	9971P	Markerville
0.84	9923P	PAS 2	6.10	1070C	Fort McMurray-Patricia McInnes
0.84	9926P	PAS 6	6.06	1057C	Genesee
0.86	9902P	AH8	5.97	9928P	PAS 8
0.86	9914P	NE10	5.94	1157C	Elk Island

0.86	1199P	Town of Bonnyville	5.76	1076C	Fort McKay South (Syncrude UE1)
0.87	9953P	South Elkton	5.76	1032C	Fort McKay-Bertha Ganter
0.87	9916P	NE7	5.75	1168C	Beaverlodge
0.87	9920P	R2	5.75	1063C	Breton
0.87	1189P	Frog Lake	5.66	1252P	St. Lina
0.88	9907P	JP102	5.38	9907P	JP102
0.88	9904P	BM11	5.26	1248C	Maskwa
0.88	9905P	BM7	5.26	1225C	Anzac
0.89	1057C	Genesee	5.04	9924P	PAS 3
0.89	9903P	BM10	5.00	1055C	Steeper
0.89	9918P	WF4	4.80	9932P	PAS 14
0.89	9909P	JP107	4.78	9909P	JP107
0.89	9915P	NE11	4.61	9925P	PAS 4
0.89	1075C	Millennium Mine	4.38	9934P	PAS 16
0.89	1092C	Caroline	4.38	9933P	PAS 15
0.90	1248C	Maskwa	4.33	9930P	PAS 12

0.90	1162C	Lamont County	4.33	9927P	PAS 7
0.90	9917P	SM8	4.31	9955P	Crossfield-Carstairs
0.90	9921P	SM7	4.31	9949P	Rimbey
0.91	1177P	Flat Lake	4.21	9916P	NE7
0.91	1168C	Beaverlodge	4.14	9922P	PAS 1
0.91	1032P	Fort McKay-Bertha Ganter	4.02	9919P	AH7
0.91	1092P	Caroline	4.00	1092C	Caroline
0.91	9955P	Crossfield-Carstairs	3.82	1250C	ST.LINA
0.91	1157C	Elk Island	3.78	9959P	Morningside
0.92	1190P	Clear Range	3.69	1190P	Clear Range
0.92	9906P	JP101	3.61	9915P	NE11
0.93	1187P	Maskwa	3.57	1071C	Fort Chipewyan
0.93	9971P	Markerville	3.46	9948P	Twin Lakes
0.93	1064C	Fort McMurray-Athabasca Valley	3.43	9962P	Grainger
0.93	1244C	Shell Muskeg River	3.41	9942P	Baseline Mountain
0.93	1244C	Shell Muskeg River	3.41	9942P	Baseline Mountain

0.93	1159C	Ross Creek	3.39	9931P	PAS 13
0.93	1191P	Fishing Lake	3.39	9923P	PAS 2
0.93	1063C	Breton	3.25	9901P	AH3
0.94	1142P	Red Deer Riverside	3.24	9941P	Sunchild
0.94	9908P	JP104	3.24	1227P	Cold Lake South Passive 2
0.94	1182P	Dupre	3.24	1193P	Cold Lake South Passive
0.94	9961P	Sunnyslope	3.22	9966P	Ferrybank
0.94	9968P	Kersey	3.18	1189P	Frog Lake
0.94	1250C	ST.LINA	2.67	9961P	Sunnyslope
0.94	1058C	Meadows	2.63	9968P	Kersey
0.94	1161C	Range Road 220	2.62	9960P	Mayton
0.94	9945P	Fallen Timber	2.62	9956P	Netook-Olds
0.94	9946P	Bearberry	2.61	9902P	AH8
0.94	9960P	Mayton	2.60	9918P	WF4
0.94	9962P	Grainger	2.59	1092P	Caroline
0.94	1054C	Carrot Creek	2.53	1195P	Fort George

0.94	1172C	Crescent Heights	2.53	1183P	La Corey
0.94	1028C	Edmonton Central	2.48	9958P	Sylvan Lake
0.94	2001C	Fort Saskatchewan-92St & 96Ave	2.43	1187P	Maskwa
0.94	1053C	Tomahawk	2.41	1192P	Beaverdam
0.94	1165C	Grande Prairie (Henry Pirker)	2.40	9954P	Bottrel
0.95	1178P	Lake Eliza	2.39	9963P	Elnora
0.95	9948P	Twin Lakes	2.37	9952P	Sundre
0.95	9949P	Rimbey	2.37	9951P	Raven River
0.95	1226C	CNRL Horizon	2.33	9965P	Samson
0.95	9959P	Morningside	2.30	1182P	Dupre
0.95	9963P	Elnora	2.30	1176P	Therien
0.95	9964P	Alix	2.21	9964P	Alix
0.95	9965P	Samson	2.00	9967P	Pakkwaw
0.95	9951P	Raven River	2.00	9950P	Leslieville
0.95	9958P	Sylvan Lake	1.96	9906P	JP101
0.96	1193P	Cold Lake South Passive	1.92	9947P	Ricinus

0.96	1227P	Cold Lake South Passive 2	1.90	1177P	Flat Lake
0.96	1029C	Edmonton East	1.88	9910P	JP205
0.96	1176P	Therien	1.87	1191P	Fishing Lake
0.96	1181P	Muriel-Kehiwin	1.85	9911P	JP210
0.96	9901P	AH3	1.76	1181P	Muriel-Kehiwin
0.96	9911P	JP210	1.76	1178P	Lake Eliza
0.96	1142C	Red Deer-Riverside	1.49	9953P	South Elkton
0.96	1174C	Cold Lake South	1.49	9946P	Bearberry
0.96	1183P	La Corey	1.21	9913P	JP213
0.96	1195P	Fort George	1.19	9970P	Panther River
0.96	9947P	Ricinus	1.19	9945P	Fallen Timber
0.96	9952P	Sundre	1.18	9914P	NE10
0.96	9950P	Leslieville	1.14	9938P	Bow Summit
0.97	1225C	Anzac	1.12	9903P	BM10
0.97	1036C	Edmonton South	1.10	9904P	BM11
0.97	1039C	Calgary Northwest	1.08	9940P	Bighorn

0.97	9956P	Netook-Olds	1.03	9921P	SM7
0.97	9967P	Pakkwaw	1.03	9917P	SM8
0.98	1070C	Fort McMurray-Patricia McInnes	0.90	9939P	Parker Ridge
0.99	1032C	Fort McKay-Bertha Ganter	0.89	9943P	Limestone Mountain
0.99	1076C	Fort McKay South (Syncrude UE1)	0.89	9905P	BM7

The spatial distribution of clusters generated at a specific level of the dissimilarity metric is another way hierarchical clustering may be used to examine the relationships between the stations. As noted earlier, this is equivalent to drawing a horizontal line across the dendrogram at a specific level of the dissimilarity metric, collecting the stations by cluster at that level, and plotting the stations on a map, colour-coded by cluster.

Figure 4.1, Figure 4.2 and Figure 4.3 show the resulting mapping for the Alberta bimonthly NO_2 clusters using the 1-R dissimilarity metric, for 1-R = 0.6, 0.5 and 0.45 (R = 0.4, 0.5 and 0.65) respectively. At 1-R dissimilarity metric levels of 0.6 and 0.5 (Figure 4.1 and Figure 4.2), the continuous monitors form a single cluster (cluster number 1), while the passive monitors form the remaining clusters. Two WCAS continuous monitors separate from the remaining continuous monitors at R=0.5 (Figure 4.2). The expanded views to the right of each of these figures for specific Airsheds show that some of the continuous monitors in WBEA and one in LICA also cluster with the passives, though the LICA passive fails to cluster with the continuous monitors by R=0.45 (Figure 4.3), and the collocated PAS continuous and passive monitors do not cluster at any of the correlation levels shown. The continuous monitors as a group remain distinct from the passive monitors until Steeper/1055C and Power/1059C form a distinct cluster with passive station St. Lina/1252P. The tendency of the continuous monitors to remain separate from the passive monitors shows that these two monitoring technologies are not providing equivalent bimonthly average results in some regions (LICA, PAS) at the given correlation levels. In the Athabasca oil sands region, passive and continuous stations located closer to the oil sands facilities tend to cluster up at 1-R = 0.5, but for levels of correlation above 0.5, the clustering between stations monitoring similar source areas is rare. Solazzo and Galmarini (2015) in their analysis of European monitoring networks found similar patterns between different European nations, noting that “The reason for this distinct country-related grouping most likely lies in the country sampling methodologies, sensitivity and data acquisition protocols not being harmonised across [the] EU”. The same is true for the Alberta passive and continuous monitoring stations – the 1-R clustering is showing that the continuous stations are more similar to each other than to the passive stations continuous and passive stations collocated at the same Airshed. This is also discussed in the previous sections of this chapter on a within-Airshed basis, for the WBEA and LICA

Airsheds. Given that collocated passive stations do not always correlate well with each other, and collocated passive stations with high correlations sometimes have high Euclidean distances (e.g. LICA 1193P and 1227P cluster at R=0.9 but have a Euclidean distance of 3.2 ppb, see Table 4.6), much of this variability seems to lie with the sampling methodology, sensitivity and data accuracy.

That the passive NO_2 monitors may suffer from sampling sensitivity issues is also suggested by the many WBEA and LICA passive monitors clustering together, despite being located in different airsheds, down to R = 0.65 (Figure 4.3, cluster “3”, red circles); the two source regions are not yet distinct at this correlation level. The converse possibility, that the temporal variability of sources and meteorology in these two regions is sufficiently similar to drive the cross-Airshed clustering, is unlikely, given the presence of other within-Airshed stations which are not part of the larger cross-Airshed cluster. Figure 3.11 shows that the passive monitors for those two regions do not become distinct until 1-R = 0.31 (R=0.69); Figure 3.12 shows that the Euclidean distances do not separate the regions into distinct clusters until Euclidean distances less than 2.77 ppbv are reached. However, Table 4.6 also shows that collocated LICA passive NO_2 monitors have Euclidean distances as great at 3.2 ppb, and Figure 3.2 and Figure 3.3 show that highly correlated and collocated (R>0.9) passive and continuous monitors (1032C/1032P) are separated by Euclidean distances of 14.4 ppb. Taken together, these findings suggest that the data from the passive monitors have sufficiently high levels of noise to make distinguishing between stations at different Airsheds difficult, and that the magnitude of the noise is as high or higher than the differences between stations within the same Airshed. One important caveat on the redundancy analysis carried out here is thus that while the stations may be ranked in order of redundancy according to the 1-R or Euclidean distances, at least some of the similarities and dissimilarities are clearly being influenced by high levels of noise within the observations.

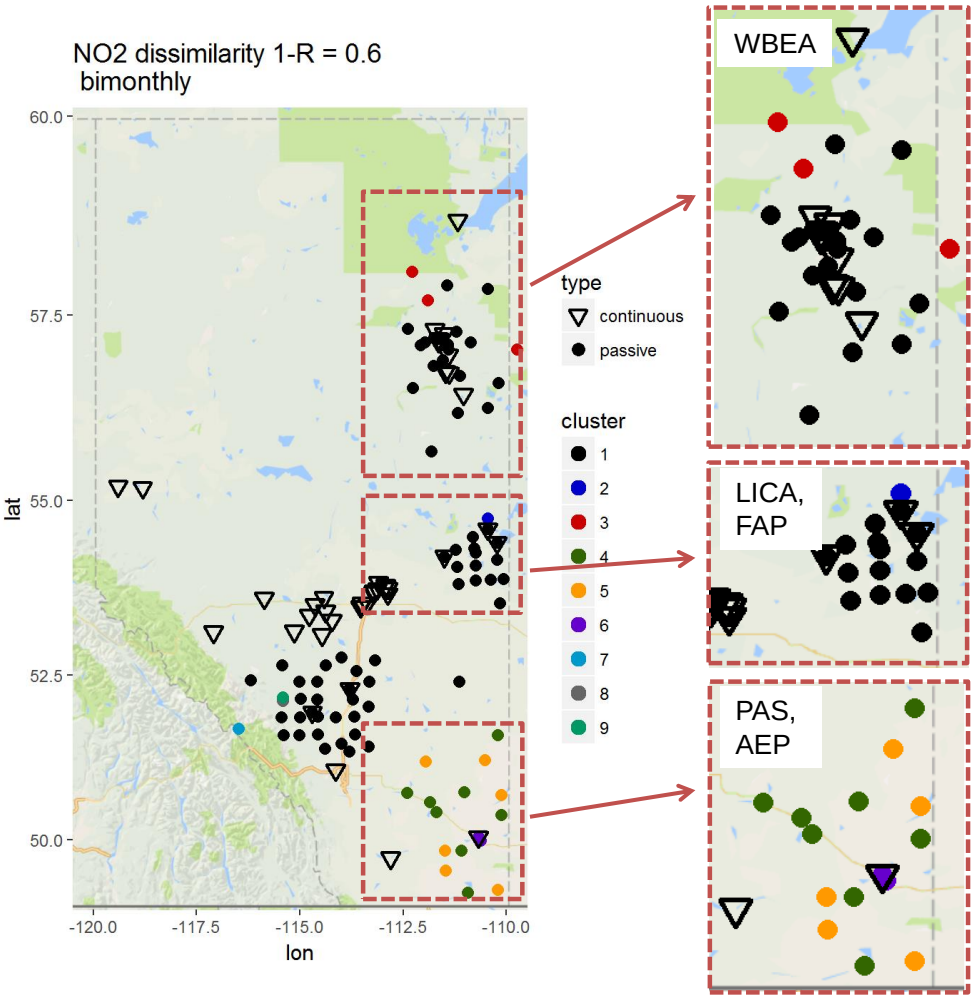


Figure 4.1 Associativity analysis for passive and continuous bimonthly NO_2 averages for 1-R = 0.6 (R=0.4) Stations are colour-coded according to cluster formation, with continuous stations are marked as triangle and passive as a circle.

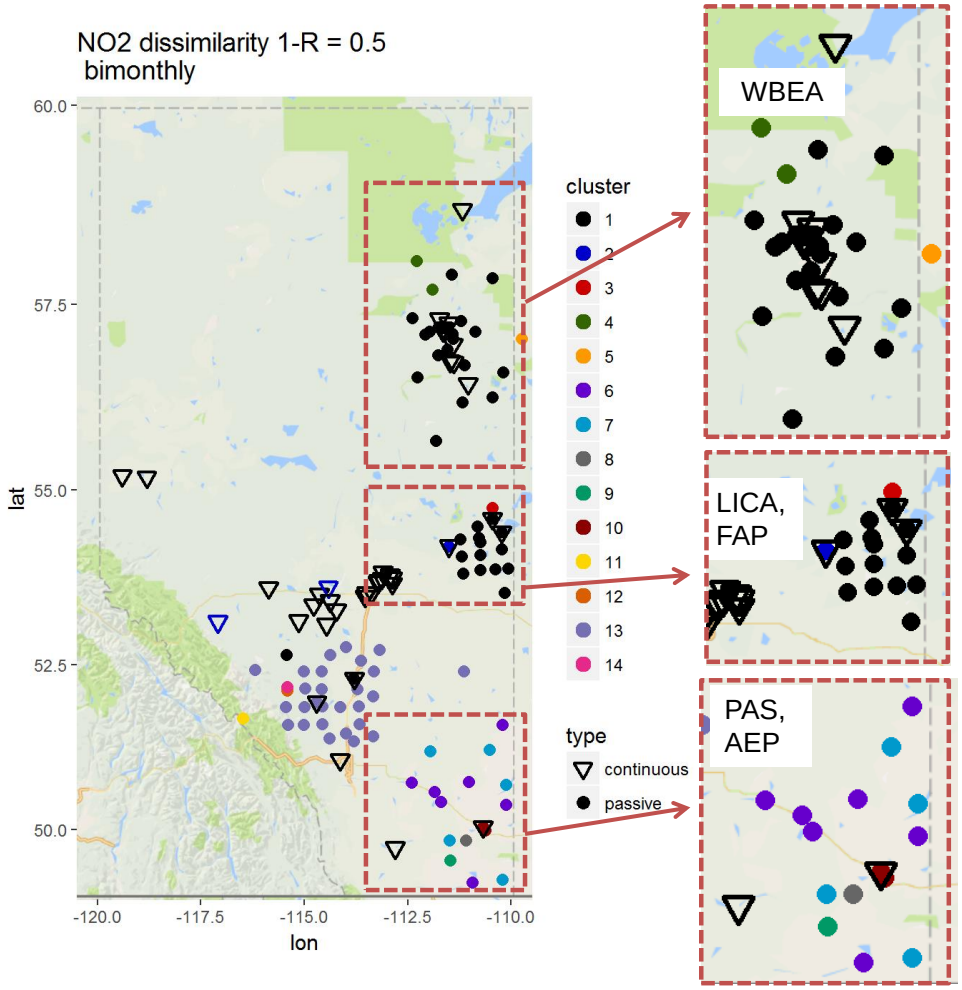


Figure 4.2 Associativity analysis for passive and continuous bimonthly NO_2 averages for 1-R = 0.5 (R=0.5) Stations are colour-coded according to cluster formation, with continuous stations are marked as triangle and passive as a circle.

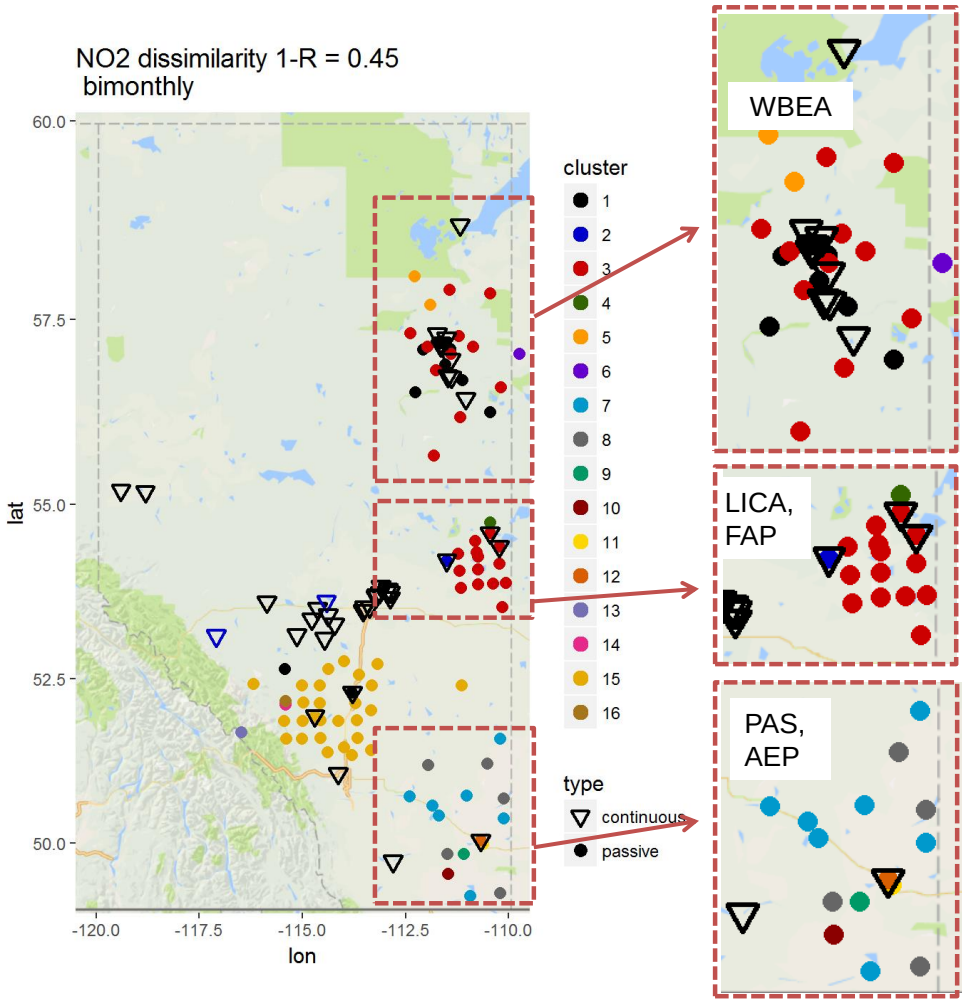


Figure 4.3 Associativity analysis for passive and continuous bimonthly NO_2 averages for 1-R = 0.45 (R=0.65) Stations are colour-coded according to cluster formation, with continuous stations are marked as triangle and passive as a circle.

The timescale filtering of the hourly continuous NO_2 data for the single year analysis may be used to show the extent to which different time scales influences the similarities (noting here that at the bimonthly averaging discussed above, all continuous monitors are part of the same 1-R cluster). Figure 4.4 shows the spatial locations of clusters occurring for 1-R = 0.6 (R=0.4), for the hourly data and time-filtered data. The hourly data (Figure 4.4 (a)) at this low level of correlation cluster across Airsheds; lower values of 1-R are required to distinguish Airsheds and local sources clearly. As variability associated with shorter time scales is removed (Figure 4.4 (b) through (d)), the number of clusters decreases, with only a single station remaining distinct at this correlation level when monthly timescales are removed (Figure 4.4 (d)). At a fixed level of correlation, the regional scale component of the NO_2 time series becomes more dominant as shorter time scales are removed. Most of the variation occurs in the shorter time scales, and all stations forming a single cluster at bimonthly time scales and lower levels of correlation is thus reasonable.

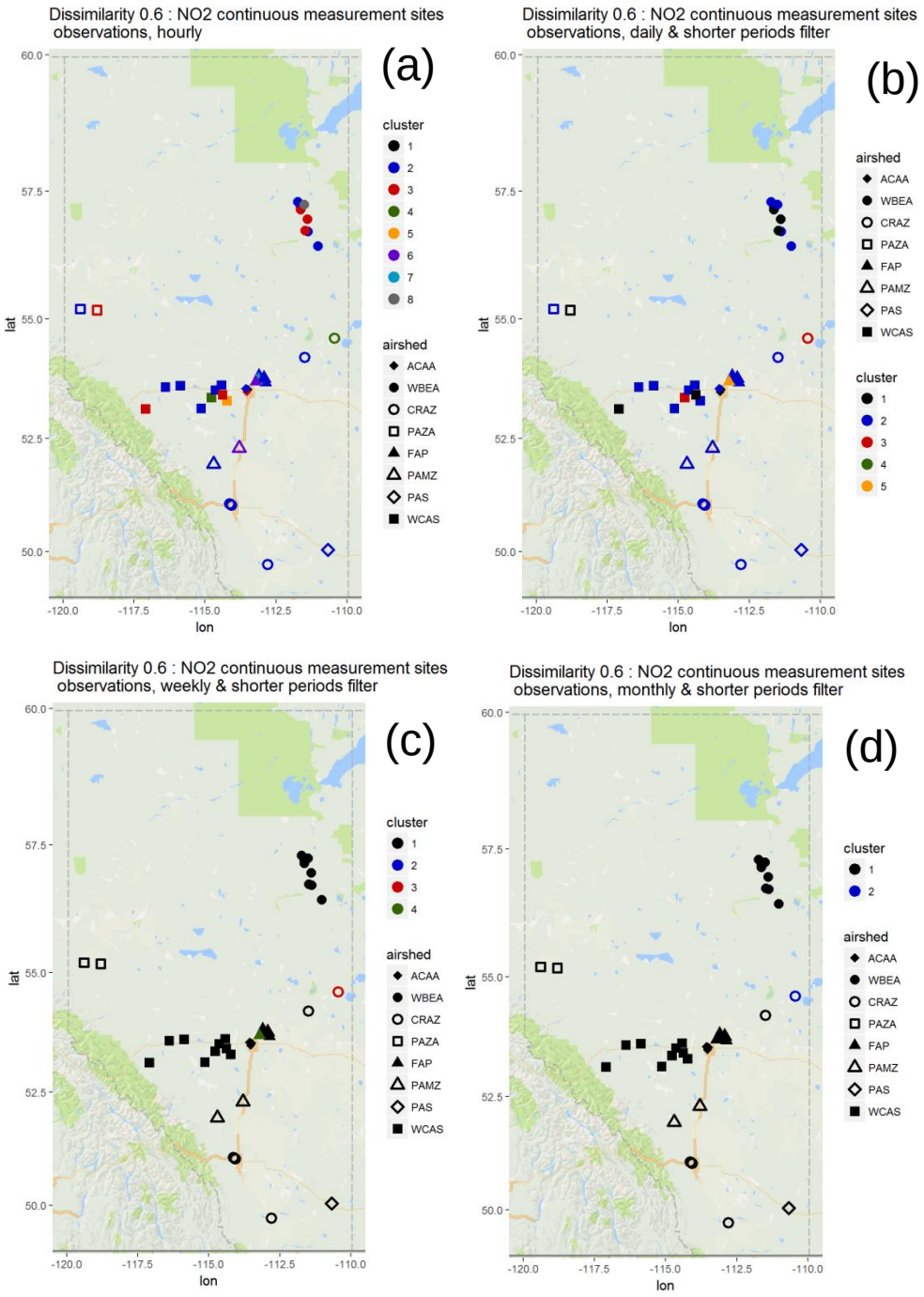


Figure 4.4 Associativity analysis for NO_2 hourly time series and the filtered time-scales using 1-R as the metric to compute the dissimilarity matrix, assuming a dissimilarity level of 0.6. Stations are colour-coded according to cluster formation, and Airsheds are plotted with different polygons.

Table 4.12 shows the 1-R and Euclidean distance dissimilarity rankings for the entire collection of Alberta continuous and passive SO_2 monitors. Once again, the ranking between the lowest 1-R (highest R) and lowest Euclidean distance stations differs significantly; e.g. only five of the stations on the left and right columns of the table for the bottom twenty rows overlap (Violet Grove, Tomahawk, and FAP stations 46,

40, and 03); stations with the highest correlations do not necessarily have the lowest Euclidean distances. However, the differences in the order of the station rankings between the two metrics are smaller than was noted for NO_2. Nevertheless, this difference in the rankings once again suggests that decisions on redundancy must thus be based on the metric which has the greater relevance towards the intended purpose of the monitoring network at these locations – long time period averages (Euclidean distance), or differences in temporal variability (1-R).

Table 4.12 Bimonthly SO_2 Similarity Ranking. Note that stations at the bottom of the two columns are the most similar (hence one measure of their level of redundancy) with respect to each metric of dissimilarity.

R	station ID	station name	EuN	station ID	station name
0.08	1156C	Redwater Industrial	24.72	1156C	Redwater Industrial
0.11	1066C	Mildred Lake	6.96	1066C	Mildred Lake
0.23	9918P	WF4	6.13	1069C	Mannix
0.25	1250C	ST.LINA	5.96	9918P	WF4
0.37	1312P	FAP-56	5.53	1250C	ST.LINA
0.43	1170C	Valleyview	4.93	9908P	JP104
0.47	1241C	Wagner2	4.23	1162C	Lamont County
0.47	1029C	Edmonton East	4.18	9907P	JP102
0.51	1198P	Hilda Lake	4.18	1074C	Lower Camp
0.53	9908P	JP104	3.32	9904P	BM11
0.55	1074C	Lower Camp	3.26	1075C	Millennium Mine
0.55	1179P	Telegraph Creek	3.26	1068C	Buffalo Viewpoint
0.55	1092C	Caroline	3.23	1244C	Shell Muskeg River
0.57	1162C	Lamont County	3.14	9917P	SM8
0.57	1059C	Power	2.99	9912P	JP212
0.57	1166C	Evergreen Park	2.99	1029C	Edmonton East
0.57	1036C	Edmonton South	2.88	9906P	JP101
0.58	9907P	JP102	2.88	9901P	AH3
0.58	1064C	Fort McMurray-Athabasca Valley	2.84	9902P	AH8

0.59	1075C	Millennium Mine	2.82	9916P	NE7
0.59	1068C	Buffalo Viewpoint	2.82	9909P	JP107
0.59	9912P	JP212	2.78	1032P	Fort McKay-Bertha Ganter
0.59	1057C	Genesee	2.60	1312P	FAP-56
0.60	1161C	Range Road 220	2.59	9915P	NE11
0.61	2001C	Fort Saskatchewan-92St & 96Ave	2.49	1241C	Wagner2
0.62	9904P	BM11	2.49	1057C	Genesee
0.62	9903P	BM10	2.48	1032C	Fort McKay-Bertha Ganter
0.62	9915P	NE11	2.48	1076C	Fort McKay South (Syncrude UE1)
0.62	1244C	Shell Muskeg River	2.34	1064C	Fort McMurray-Athabasca Valley
0.64	1159C	Ross Creek	2.32	1070C	Fort McMurray-Patricia McInnes
0.64	1248C	Maskwa	2.32	1226C	CNRL Horizon
0.65	1069C	Mannix	2.28	1248C	Maskwa
0.66	1032P	Fort McKay-Bertha Ganter	2.25	1198P	Hilda Lake
0.66	1226C	CNRL Horizon	2.11	9913P	JP213
0.67	2000C	Bruderheim	2.08	1159C	Ross Creek
0.71	9902P	AH8	2.05	9910P	JP205
0.71	1292P	FAP-35	1.99	1161C	Range Road 220
0.72	1063C	Breton	1.99	2000C	Bruderheim
0.72	1070C	Fort McMurray-Patricia McInnes	1.96	1049C	Lethbridge
0.73	1157C	Elk Island	1.95	1170C	Valleyview
0.73	1174C	Cold Lake South	1.95	1092C	Caroline
0.73	9909P	JP107	1.94	1187P	Maskwa
0.75	1272P	FAP-15	1.89	1292P	FAP-35
0.75	1049C	Lethbridge	1.87	9903P	BM10

0.75	1172C	Crescent Heights	1.81	1174C	Cold Lake South
0.75	1225C	Anzac	1.78	1142C	Red Deer-Riverside
0.76	1032C	Fort McKay-Bertha Ganter	1.78	1165C	Grande Prairie (Henry Pirker)
0.76	1076C	Fort McKay South (Syncrude UE1)	1.74	1036C	Edmonton South
0.77	1062C	Edson	1.73	9911P	JP210
0.78	1199P	Town of Bonnyville	1.73	1225C	Anzac
0.80	9901P	AH3	1.68	1058C	Meadows
0.81	1186P	Primrose	1.56	1252P	St. Lina
0.81	1187P	Maskwa	1.56	1197P	Mahihkan
0.82	1259P	FAP-02	1.54	2001C	Fort Saskatchewan-92St & 96Ave
0.82	9913P	JP213	1.47	1277P	FAP-20
0.83	1167C	Smoky Heights	1.47	1259P	FAP-02
0.83	1168C	Beaverlodge	1.41	1303P	FAP-47
0.84	1306P	FAP-50	1.40	1186P	Primrose
0.85	9911P	JP210	1.40	1157C	Elk Island
0.85	9906P	JP101	1.34	1059C	Power
0.85	1176P	Therien	1.28	9914P	NE10
0.85	1252P	St. Lina	1.28	9905P	BM7
0.85	1285P	FAP-28	1.27	1199P	Town of Bonnyville
0.85	1279P	FAP-22	1.27	1291P	FAP-34
0.86	9914P	NE10	1.22	1272P	FAP-15
0.86	1052C	Violet Grove	1.22	1179P	Telegraph Creek
0.86	1053C	Tomahawk	1.19	1071C	Fort Chipewyan
0.86	1192P	Beaverdam	1.19	1054C	Carrot Creek
0.87	1058C	Meadows	1.15	1168C	Beaverlodge
0.87	9916P	NE7	1.12	1167C	Smoky Heights

0.87	9910P	JP205	1.12	1062C	Edson
0.88	1197P	Mahihkan	1.08	1052C	Violet Grove
0.88	1183P	La Corey	1.08	1053C	Tomahawk
0.88	1303P	FAP-47	1.00	1166C	Evergreen Park
0.88	1302P	FAP-46	1.00	1063C	Breton
0.88	1297P	FAP-40	0.98	1302P	FAP-46
0.88	1260P	FAP-03	0.98	1278P	FAP-21
0.88	1189P	Frog Lake	0.92	1055C	Steeper
0.88	1182P	Dupre	0.92	1172C	Crescent Heights
0.89	1311P	FAP-55	0.88	1297P	FAP-40
0.89	1262P	FAP-05	0.88	1260P	FAP-03
0.89	9905P	BM7	0.86	1176P	Therien
0.89	1178P	Lake Eliza	0.83	1285P	FAP-28
0.89	1190P	Clear Range	0.83	1279P	FAP-22
0.91	1278P	FAP-21	0.82	1189P	Frog Lake
0.91	1277P	FAP-20	0.82	1182P	Dupre
0.91	1227P	Cold Lake South Passive 2	0.82	1306P	FAP-50
0.91	1193P	Cold Lake South Passive	0.80	1177P	Flat Lake
0.91	1195P	Fort George	0.79	1311P	FAP-55
0.91	1177P	Flat Lake	0.79	1262P	FAP-05
0.92	1291P	FAP-34	0.73	1178P	Lake Eliza
0.92	1054C	Carrot Creek	0.73	1190P	Clear Range
0.93	9917P	SM8	0.70	1195P	Fort George
0.93	1071C	Fort Chipewyan	0.70	1192P	Beaverdam
0.93	1055C	Steeper	0.61	1183P	La Corey
0.97	1142C	Red Deer-Riverside	0.50	1227P	Cold Lake South Passive 2

0.97	1165C	Grande Prairie (Henry Pirker)	0.50	1193P	Cold Lake South Passive

The Alberta bimonthly SO_2 dendrogram was used to generate clusters at 1-R values of 0.8, 0.7 and 0.6 (R = 0.2 0.3, and 0.4, respectively for spatial mapping (Figure 4.5 (a),(b),(c)). The SO_2 clusters are considerably more discontinuous (more clusters for a given area and correlation level) than NO_2, showing the influence of more discrete local sources and conditions as opposed to the regional influences apparent for NO_2 at low correlation levels (compare Figure 4.5 (a) continuous monitors those of with Figure 4.1). At low correlation levels, many of the SO_2 stations, both passive and continuous, fall into the same cluster (Figure 4.5(a), cluster 3), and widely separated stations fall into the same cluster (e.g. some FAP stations cluster with WBEA stations (Figure 4.5 (a), cluster 1). These anomalies persist to higher correlation levels (Figure 4.5 (c), cluster 4 including passive stations in WBEA, FAP and LICA, along with WCAS and LICA continuous stations. At the highest correlation level shown here (R= 0.4), collocated continuous and passive stations in LICA and PAS do not form clusters, and many stations which are located in different areas of the province fall within the same cluster (Figure 4.5 (c)). As noted earlier, NO_2 and SO_2 are influenced by different source types (primarily large stack sources for SO_2, and surface area mobile sources for NO_2), and these differences help explain a more spatially discontinuous pattern of clusters (at a given 1-R level) for SO_2 than for NO_2. However, the 1-R clustering of passive stations and some continuous stations across large spatial distance suggests that the SO_2 data are similar for reasons which at present are unknown, but may include sources with similar time signatures to their emissions, similarly *low* concentrations, or the same issues as noted in the NO_2 analysis (sampling methodology, instrument sensitivity and/or data acquisition protocols not being harmonized, and the likelihood, based on collocated passive and continuous monitors, of a high degree of error within the measurements).

The one year dataset of hourly continuous SO_2 observations were analysed in the same fashion as for those NO_2 to show spatial relationships of the clusters, with the results shown in Figure 4.6, at a 1-R dissimilarity level of 0.7 (R=0.3). Even at this low level of correlation, the clustering of the unfiltered hourly data (Figure 4.6 (a)) shows a large degree of spatial variability (large number of clusters), reflecting the spatial and temporal heterogeneity of SO2 sources.

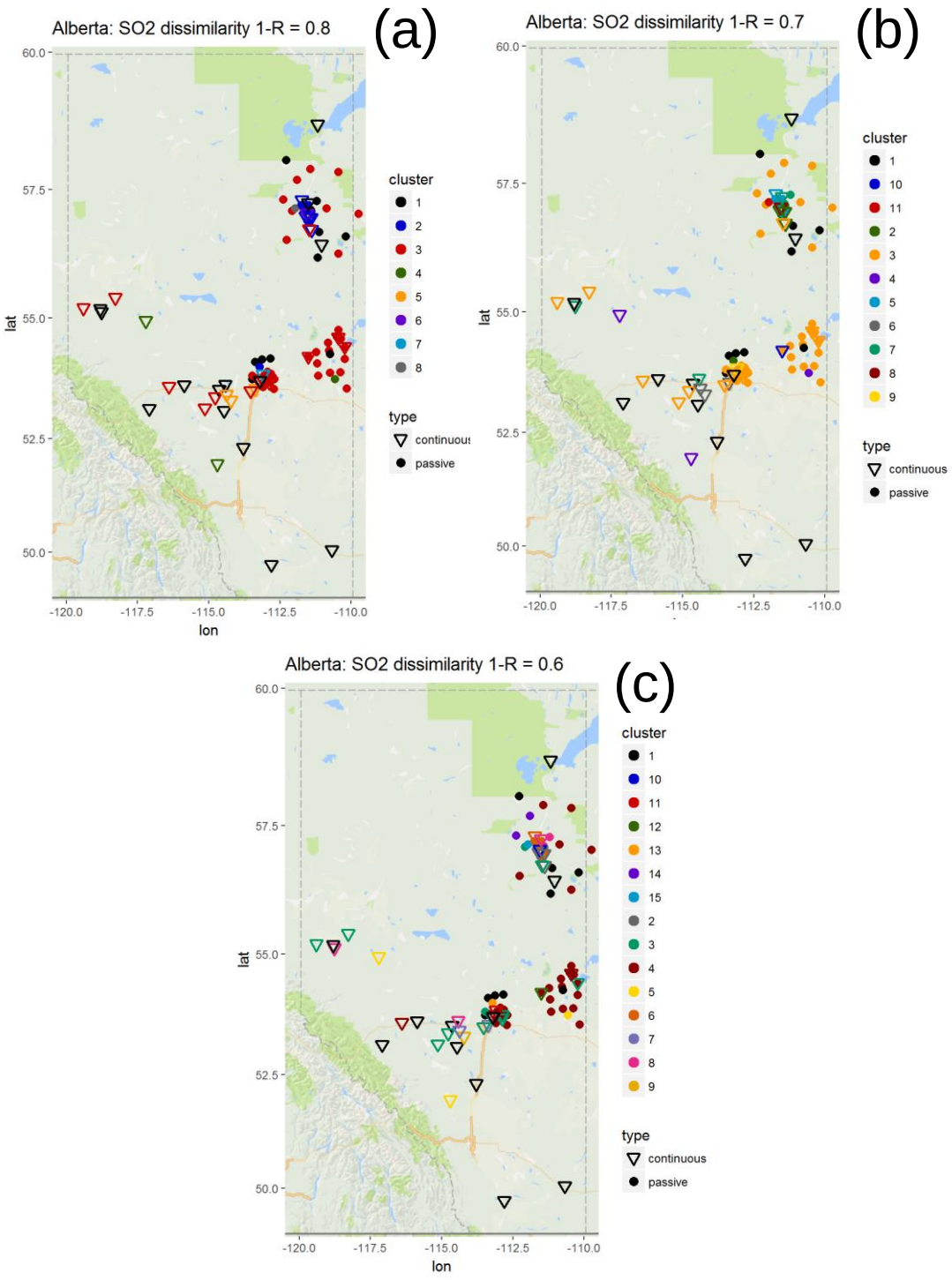


Figure 4.5 Associativity analysis for passive and continuous bimonthly SO_2 averages using 1-R as the dissimilarity metric, assuming a dissimilarity level of 0.7, 0.6 and 0.5 (R=0.3, 0.4, 0.5). Stations are colour-coded according to cluster formation, with continuous stations are marked as triangle and passive as a circle.

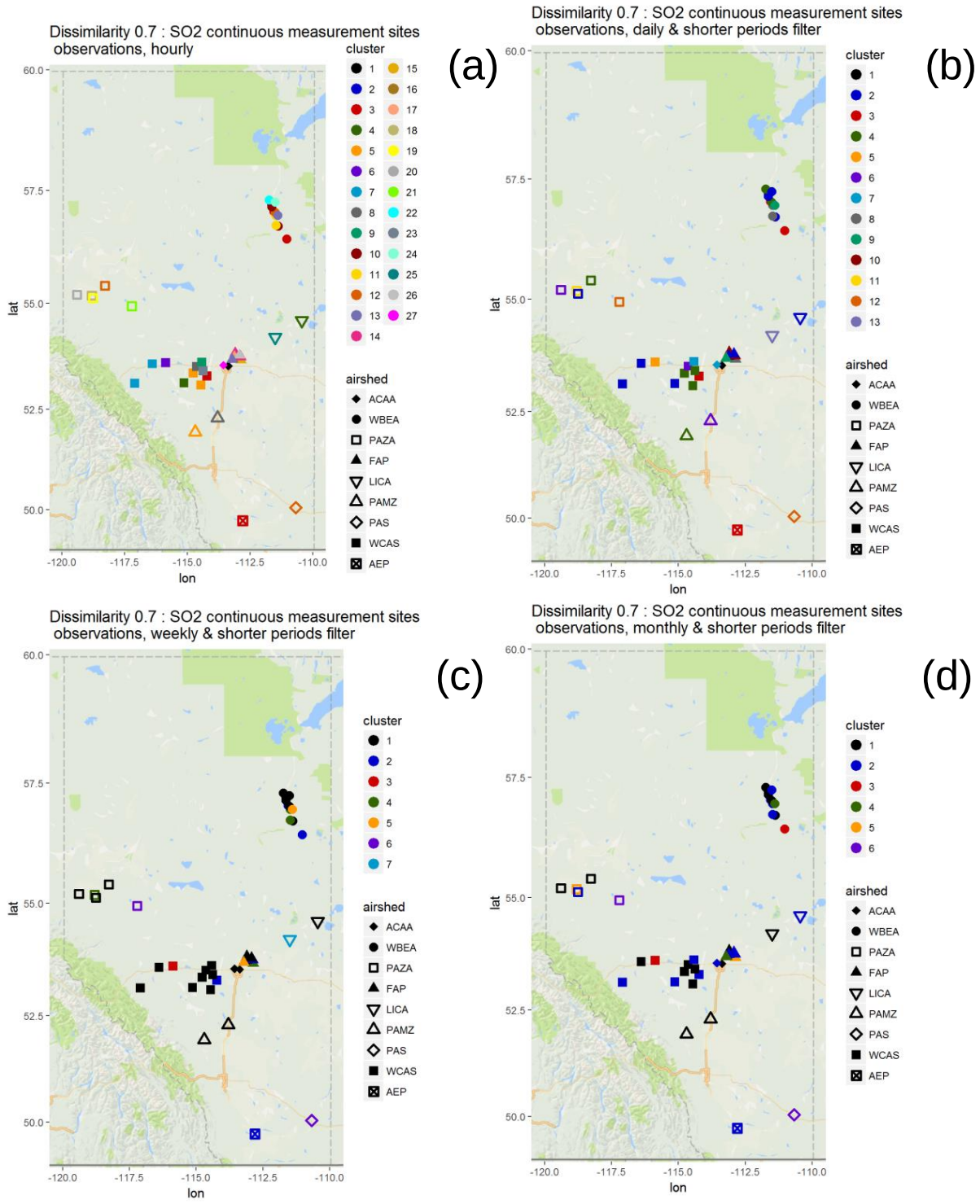


Figure 4.6 Associativity analysis for SO_2 hourly time series and the filtered time-scales considering 1-R as the metric to compute the dissimilarity matrix, assuming a dissimilarity level of 0.7. Stations are colour-codded according to cluster formation, and Airsheds are plotted with different polygons.

In order to better discuss possible redundancies for SO_2 the 1-R clusters (Figure 3.13) which contain four or more stations were collected in Table 4.13. The clusters are ordered in terms of increasing R. The maximum and minimum Euclidean distance between stations within each of these 1-R clusters is shown in the final two columns of the Table (the Euclidean distances were determined by finding the largest Euclidean distance between pairs of stations in the 1-R cluster, using the Euclidean distance dendrogram, Figure 3.14). The following observations may be made, based on Table 4.13:

1) The clusters are mainly represented by WBEA stations (cluster numbers 1, 3, 4 5, and 6, 7 and 8) with a correlation level ranging from 0.53 to 0.80. There are also three clusters for FAP (cluster numbers 1, 2 and 5) with correlation ranging from 0.67 to 0.84, and a single cluster for LICA
2) Clusters comprised of WBEA stations tend to have lower correlations but higher maximum Euclidean distance values (2.82 to 6.13 ppbv) than other Airsheds, suggesting high source emissions and greater variability compared to the other clusters.
3) FAP clusters have lower variability between correlation and the difference between the lowest and highest Euclidean distance value is the smallest, suggesting many of these monitors are sampling similar sources or background concentrations.
4) LICA monitors are having a difference of ~1ppbv between the lowest and the highest Euclidean distance values and as a reasonable correlation (0.75). Though spread out spatially, these passives monitors are measuring similar sources or background concentrations.

Table 4.13 and the associated discussion thus provides a methodology whereby both sets of clustering information may be combined to rank stations, providing information on relative levels of similarity and hence potential redundancies, with regards to both metrics.

Table 4.13 1-R SO_2 clusters selected from Figure 3.13, with the largest Euclidean distance between their members (from Figure 3.14).

1-R Cluster Number (Airshed Abbreviation)	1-R Cluster Members	1-R	R	Euclidean Distance between 1-R members (ppbv)	
				largest	smallest
1(FAP)	Brudenheim (C) FAP-15 (P) FAP-22 (P) FAP-28 (P) FAP-03 (P) FAP-40 (P) FAP-46 (P)	0.33	0.67	1.99	0.83

2(FAP)	FAP-02 (P) FAP-20 (P) FAP-21 (P) FAP-47 (P)	0.18	0.82	1.47	0.98
3(LICA)	Cold Lake South Passive (P) Cold Lake South Passive 2 (P) Dupre (P) Frog Lake (P) La Corey (P) Mahihkan (P) Beaversdam (P) Flat Lake (P) Fort George (P) Clear Range (P) Lake Eliza (P) St. Lina (P) Therien (P)	0.25	0.75	1.56	0.50
3(WBEA)	JP107 (P) JP213 (P) JP205 (P) NE7 (P)	0.27	0.73	2.82	2.05
4(WBEA)	JP104 (P) Shell Muskeg River (C) NE11 (P)	0.47	0.53	4.93	2.59
5(FAP)	FAP-50 (P) FAP-05 (P) FAP-55 (P)	0.16	0.84	8.16	7.9
6(WBEA)	AH3 (P) NE10 (P) BM7 (P) Fort Chipewyan (C) SM8 (P)	0.20	0.80	3.13	1.19
7(WBEA)	Buffalo Viewpoint (C) Millenium Mine (C) Lower Camp (C) Fort McMurray-Athabasca Valley (C) JP102 (P)	0.57	0.43	4.18	2.34

8(WBEA)	Fort Mackay-Bertha South (C) Fort Mackay-Bertha Ganter (C) Mannix (C) CNRL Horizon (C) Fort Mackay-Bertha Ganter (P)	0.36	0.64	6.13	2.13

4.1.4 All Alberta: Continuous Monitoring for Multiple Chemical Species

Figures 3.15 and 3.16 illustrate some of the complexities in determining potential redundancies for the suites of continuous monitoring stations. For example, all WBEA continuous O_3 monitors are highly correlated and cluster together as successively longer time scales being removed with respect to correlation (Figure 3.15(a) through (c)), but maintain that clustering from the standpoint of magnitude only for original hourly and daily timescales (Figure 3.16(a),(b)). The correlation coefficients all increase as the timescale of the variations being removed is increased. The Euclidean distances also decrease as successively longer timescales are removed –the stations become more similar to each other, implying that most of the variability with regards to 1-R and Eulerian distance resides at shorter timescales. The residual signal in the data records at longer timescales is highly similar, implying a greater degree of overall potential redundancy with increasing timescale. The ozone figures thus provide a good example of one of the central themes of the continuous monitor analysis – the way the desired purpose of the monitoring has bearing on the assessment of redundancy. If the component of ozone concentrations which constitutes the long-term regional background signal (which is isolated by the filtering out of time scales smaller than a month) is a key result of the continuous ozone monitoring network, many of these monitors might be interpreted as redundant based on this similarity, since many will have similar Euclidean distances and high correlation coefficients. However, if the shorter timescale variations are important, then far fewer of the stations might be interpreted as redundant, since at shorter timescales the correlation levels are lower and the Euclidean distances are higher, even within a given Airshed.

Given an a priori decision on the relevant timescale and metric for redundancy, the colour-coded station lists on the right hand side of each of the panels the Figure 3.15 through 3.34 may be used to assess potential relative redundancy based on similarity, for the metrics examined here. The station names appearing at the bottom of the lists are the most similar, hence potentially the most redundant; the station names at the top of the list are the least redundant.

An example for unfiltered hourly data and the 1-R metric, the stations with the four highest correlation coefficients, and the stations with the four lowest correlation coefficients, are shown in Table 4.14. It should be noted that the “highest” values within the upper half of Table 4.14 refer to the highest correlations for the given species, within that species. The correlations themselves are not always particularly high in an absolute sense, and depend on the species under consideration. Ranked in order from highest to lowest, two groups can be seen, with higher values for NO_x, NO, NO_2, O_3 and $PM_{2.5}$, and lower values for NMHC, CH_4, THC and TRS. This split may represent differences in the relative accuracy of the sampling methodology in each case, the dynamic range of the chemical being measured (e.g. CH_4 has a high “background” concentration and this may affect the correlations), and/or differences in the

locations and variety of sources between the different chemicals. As noted throughout this work, the methodology provides a relative ranking; “more” or “less” similar for a given metric, hence potentially more or less redundant, but a single number of 1-R thus cannot be used to represent a limit for redundancy for all species. The stations in in the upper part of Table 4.14 are the more similar, for hourly timescales, using 1-R as the metric.

Table 4.15 is a similar table for the hourly stations for the Euclidean distance metric of dissimilarity. Larger tables could be reconstructed from the hourly (or other timescale) dendrograms of Figures 3.15 to 3.34). Table 4.14, Table 4.15, and extensions of these tables using the station rankings appearing to the right of the dendrograms in Figures 3.15 to 3.34, may be used to assign relative redundancy levels for the two metrics examined here, and may also be used to determine the potential penalty in assessing a given station as potentially redundant. For example, for the chemical NO, and the 1-R dissimilarity metric (Table 4.14), the two most similar stations are Fort McKay South/1076 and Fort McKay Bertha Ganter/1032. If one of those stations were to be removed, then the other station would be expected to correlate with the missing station to a level of R=0.81. However, these two stations’ Euclidean distances from each other on an hourly basis is 657 ppb from Figure 3.20, compared to the lowest Euclidean distance for hourly NO being 80.5 ppb, indicating that at times elevated NO concentrations are measured at one station but not at the other. The two stations are 4 km apart, with an elevation difference of 6 metres, and are located in a shallow river valley. A comparison of the time series for the two stations shows that many of the events with elevated NO concentrations at the two sites coincide, and last several hours, though the arrival times of the peak concentrations are often offset by an hour, and occasional peaks occur which are much higher in one station compared the other. The stations are located to the north and south of a settlement located in a shallow river valley, and may be exposed to high concentration plumes from different emissions sources (which may also reach one station but not the other); this sequence of events seems plausible given the data record and geographical information, and explains the high correlation yet high Euclidean distance resulting from the clustering analysis. The Euclidean distances for NO decrease between these and other stations as shorter time scales are removed (e.g. 80.4 ppb when monthly and shorter time scales are filtered out) – the stations become more redundant in the residual “background” concentration signal, but are much less redundant from the standpoint of shorter term high concentration events.

The time scale of interest for monitoring must therefore be taken into consideration in determining potential redundancy based on the similarity analysis. The above illustration describes the process by which the analysis done here may better inform decision making, but those decisions clearly must be made on the basis of individual chemical species, and must take into account the time scale(s) of interest for the monitoring network, and other reasons for the placement of the monitors.

Table 4.14 Stations with the four highest, four lowest correlation coefficients, hourly observations, by chemical species.

Highest Correlation Coefficient Stations, R (most similar to other stations)		
NO 0.8105 (1076) Fort McKay South (Syncrude UE1) 0.8106 (1032) Fort McKay-Bertha Ganter 0.7786 (1157) Elk Island 0.7786 (1162) Lamont County	**NO_2** 0.8899 (2001) Fort Saskatchewan - 92 St and 96 Ave 0.8899 (1159) Ross Creek 0.8539 (2002) Woodcroft 0.8539 (1028) Edmonton Central	**NO_x** 0.8403 (1157) Elk Island 0.8403 (1162) Lamont County 0.8393 (1032) Fort McKay-Bertha Ganter 0.8393 (1076) Fort McKay South (Syncrude UE1)
O_3 0.9316 (1032) Fort McKay-Bertha Ganter 0.9316 (1076) Fort McKay South (Syncrude UE1) 0.9286 (2002) Woodcroft 0.9286 (1036) Edmonton Central	**$PM_{2.5}$** 0.8218 (1032) Fort McKay-Bertha Ganter 0.8218 (1076) Fort McKay South (Syncrude UE1) 0.7623 (1221) Calgary Central 2 0.7623 (1039) Calgary Northwest	**SO_2** 0.8188 (1032) Fort McKay-Bertha Ganter 0.8188 (1076) Fort McKay South (Syncrude UE1) 0.6570 (1070) Fort McMurray-Patricia McInnis 0.6570 (1064) Fort McMurray-Athabasca Valley
CH_4 0.6038 (1221) Calgary Central 2 0.6038 (1039) Calgary Northwest 0.5343 (1070) Fort McMurray-Patricia McInnis 0.5343 (1064) Fort McMurray-Athabasca Valley	**NMHC** 0.3744 (1221) Calgary Central 2 0.3744 (1039) Calgary Northwest 0.2457 (1070) Fort McMurray-Patricia McInnis 0.2457 (1064) Fort McMurray-Athabasca Valley	**THC** 0.8057 (1032) Fort McKay-Bertha Ganter 0.8057 (1076) Fort McKay South (Syncrude UE1) 0.6232 (1028) Edmonton Central 0.6232 (1036) Edmonton South
TRS 0.7234 (1032) Fort McKay-Bertha Ganter 0.7234 (1076) Fort McKay South (Syncrude UE1) 0.5513 (1072) Barge Landing 0.3684 (1165) Grande Prairie (Henry Pirker), (1166) Evergreen Park		
Lowest Correlation Coefficient Stations, R (least similar to other stations)		
NO 0.0369 (1055) Steeper 0.0664 (1225) Anzac 0.2045 (1248) Maskwa 0.2062 (1064) Fort McMurray-Athabasca Valley	**NO_2** 0.2782 (1248) Maskwa 0.3885 (1225) Anzac 0.4336 (1057) Genesee 0.4336 (1241) Wagner2	**NO_x** 0.2580 (1248) Maskwa 0.2607 (1225) Anzac 0.3171 (1064) Fort McMurray-Athabasca Valley 0.3525 (1172) Crescent Heights
O_3	**$PM_{2.5}$**	**SO_2**

0.3475 (1055) Steeper 0.6110 (1057) Genesee 0.6662 (1168) Beaverlodge 0.6662 (1165) Grande Prairie (Henry Pirker)	0.1881 (1056) Hinton 0.3212 (1049) Lethbridge 0.4149 (1156) Redwater Industrial 0.4487 (1055) Steeper	0.0406 (1092) Caroline 0.0406 (1156) Redwater Industrial 0.1237 (1167) Smoky Heights 0.2137 (1170) Valleyview
CH_4 0.3057 (1162) Lamont County 0.3971 (1142) Red Deer-Riverside 0.4004 (1225) Anzac 0.4682 (1032) Fort McKay-Bertha Ganter	**NMHC** 0.0905 (1161) Range Road 220 0.0910 (1225) Anzac 0.0910 (1225) Anzac 0.1680 (2000) Bruderheim	**THC** 0.2185 (1248) Maskwa 0.2395 (1250) St. Lina 0.2420 (1165) Grande Prairie (Henry Pirker) 0.3362 (1029) Edmonton East
TRS 0.0427 (1092) Caroline 0.0470 (1167) Smoky Heights 0.0906 (1225) Anzac 0.1977 (1056) Hinton		

Table 4.15 Stations with the four lowest and four highest Euclidean distances.

Highest Correlation Coefficient Stations, R (most similar to other stations)		
NO $8.05x10^1$ (1250) St. Lina $8.05x10^1$ (1092) Caroline $8.40x10^1$ (1055) Steeper $1.07x10^2$ (1059) Power	**NO_2** $2.69x10^2$ (1162) Lamont County $2.69x10^2$ (1157) Elk Island $2.82x10^2$ (1055) Steeper $2.82x10^2$ (1250) St. Lina	**NO_x** $3.24x10^2$ (1055) Steeper $3.24x10^2$ (1250) St. Lina $3.41x10^2$ (1162) Lamont County $3.41x10^2$ (1157) Elk Island
O_3 0.470 (1076) Fort McKay South 0.470 (1032) Fort McKay-Bertha Ganter 0.501 (1029) Edmonton East 0.501 (1028) Edmonton Central	**$PM_{2.5}$** $2.89x10^2$ (1059) Power $2.89x10^2$ (1053) Tomahawk $2.98x10^2$ (1057) Genesee $4.02x10^2$ (1062) Edson	**SO_2** $2.50x10^1$ (1055) Steeper $2.50x10^1$ (1142) Red-Deer-Riverside $4.56x10^1$ (1166) Evergreen Park $5.23x10^1$ (1172) Crescent Heights
CH_4 9.07 (1070) Fort McMurray-Patricia McInnes	**NMHC** $4.37x10^3$ (1064) Fort McMurray-Athabasca Valley	**THC** $1.09x10^1$ (1070) Fort McMurray-Patricia McInnes

9.07 (1064) Fort McMurray-Athabasca Valley 9.96 (1039) Calgary Nortwest 9.96 (1221) Calgary Central 2	$4.37x10^3$ (1225) Anzac $4.50x10^3$ (1028) Edmonton Central $4.50x10^3$ (1039) Calgary Northwest	$1.09x10^1$ (1064) Fort McMurray-Athabasca Valley $1.28x10^1$ (1221) Calgary Central $1.28x10^1$ (1039) Calgary Northwest
TRS $2.66x10^1$ (1167) Smoky Heights $2.66x10^1$ (1092) Caroline $2.82x10^1$ (1076) Fort McKay South $2.82x10^1$ (1032) Fort McKay-Bertha Ganter		
Lowest Correlation Coefficient Stations, R (least similar to other stations)		
NO $2.02x10^3$ (1156) Redwater Industrial $1.65x10^3$ (1221) Calgary Central 2 $1.45x10^3$ (1244) Shell Muskeg River $1.45x10^3$ (1165) Grande Prairie (Henry Pirker)	**NO_2** $1.05x10^3$ (1244) Shell Muskeg River $9.84x10^2$ (1064) Fort McMurray-Athabasca Valley $9.84x10^2$ (1075) Millennium Mine $9.41x10^2$ (1165) Grande Prairie (Henry Pirker)	**NO_x** $2.84x10^3$ (1156) Redwater Industrial $2.71x10^3$ (1221) Calgary Central 2 $2.24x10^3$ (1244) Shell Muskeg River $2.12x10^3$ (1165) Grande Prairie (Henry Pirker)
O_3 1.04 (1057) Genesee 0.905 (1172) Crescent Heights 0.905 (1049) Lethbridge 0.888 (1168) Beaverlodge	**$PM_{2.5}$** $8.69x10^2$ (1167) Smoky Heights $8.36x10^2$ (1226) CNRL Horizon $8.36x10^2$ (1244) Shell Muskeg River $8.03x10^2$ (1028) Edmonton Central	**SO_2** $4.89x10^2$ (1075) Millennium Mine $4.40x10^2$ (1244) Shell Muskeg River $3.97x10^2$ (1074) Lower Camp $2.99x10^2$ (1226) CNRL Horizon
CH_4 $2.59x10^1$ (1142) Red Deer-Riverside $1.93x10^1$ (1029) Edmonton East $1.84x10^1$ (1028) Edmonton Central $1.84x10^1$ (1161) Range Road 220	**NMHC** $1.96x10^4$ (2000) Bruderheim $1.58x10^4$ (1161) Range Road 220 $8.26x10^3$ (1032) Fort McKay-Bertha Ganter $7.11x10^3$ (1221) Calgary Central 2	**THC** $5.03x10^1$ (2000) Bruderheim $4.31x10^1$ (1142) Red-Deer-Riverside $3.99x10^1$ (1092) Caroline $3.91x10^1$ (1244) Shell Muskeg River
TRS $5.91x10^1$ (1165) Grande Prairie (Henry Pirker) $5.49x10^1$ (1075) Millennium Mine $5.13x10^1$ (1225) Anzac $4.63x10^1$ (1064) Fort McMurray-Athabasca Valley		

4.2 Comparisons of Hierarchical Clustering Results using Time-filtered versus Time-averaged Data

In the analysis described in Section 4.1.4, it was noted that as successively larger time scales are filtered from the data used for clustering. Analyzing the dendograms, the magnitudes of the clustering metrics show an increasingly higher degree of similarity, though the stations no longer cluster according to Airshed. This is shown by the tendency of the clusters to move nearer to the x-axis in the dendrograms for both metrics as the time-filtering removed successively longer timescales, and for the clusters to no longer group according the same colour-coded Airshed (most noticeable for the 1-R metric). These findings indicate that much of the correlation signal may be found on the shorter timescales *within* individual Airsheds, and on the longer timescales, the regional background residual correlation signal becomes more similar *across* Airsheds. The overall decrease in Eulerian metric values as increasingly larger time scales are removed indicates that the residual concentrations are also becoming similar in magnitude. This effect varied depending on the chemical species, and was stronger in those species for which short-term "spikes" in concentration are surrounded by lower concentration background levels (such as SO_2) to those with a smaller dynamic range in concentration (such as O_3; compare 1-R dendrograms at different time scales for SO_2 (Figure 3.23) and O_3 (Figure 3.15).

While these results demonstrate that most of the concentration signal which identifies specific Airsheds as unique resides in the shorter time scales, this does not mean that this information is necessarily lost for observations that comprise long-term averages. The KZ time-filtering removes the information in short time scales, but observations which are averaged across time periods do not necessarily lose this information, since the high frequency signal is incorporated in the average. The relative impact of time-filtering versus averaging is explored in this section, though carrying out hierarchical clustering for data sets which are time-filtered and averaged, and comparing the resulting dendrograms.

Here we show the results for hierarchical clustering analysis for NO_2 and SO_2 when 1-year continuous observations were *averaged* daily (365 values), weekly (52 values) and monthly (12 values), and are compared to the dendrograms in which the daily, weekly and monthly time-scales have been *removed* by KZ filtering. The results from these tests for NO2 are compared in Figure 4.7 (1-R) and Figure 4.8 (Euclidean distance), with Figure 4.9 and Figure 4.10 providing the equivalent comparison for SO2.

The results for NO_2 using the 1-R metric (Figure 4.7) show that different clusters are being generated at all levels of averaging or time-filtering (compare columns of panels in the Figure; the pattern of clustering changes), and these differences become more pronounced for the longer time intervals of either averaging or scaling. Clearly different information is being retained via the two processes. However, at longer time averages and scales (Figure 4.7 (c),(f),), monthly time scales removed (c) and averaged (f) most of the tendency of the data to cluster within Airsheds has been lost, aside from the WBEA stations. For the corresponding Euclidean distance analysis (Figure 4.8), the differences between averaging and time-filtering is even more pronounced.

The results for SO_2 using 1-R (Figure 4.9) and Euclidean distance (Figure 4.10) as the metric to compute dissilarity show an even larger difference between clusters generated using time-filtered data (top row of panels) and time-averaged data (bottom row of panels) than NO_2. As noted earlier, the SO_2 time series

are more likely to be composed of short-term “spikes” in concentration surrounded by lower “background” levels than NO_2. The short-term spikes will be included in time averaging but removed in time-filtering, driving the larger differences between SO_2 than NO_2 dendrograms.

The averaging results (bottom row of panels in in Figure 4.7, Figure 4.8, Figure 4.9, Figure 4.10) also show that averaging loses some of the information identifying an airshed and its sources as unique. That is, comparing panels (e – weekly averages) and (f – monthly averages) in these figures, the pattern of clustering changes, and there is a reduced tendency for clusters of stations to be found within Airsheds. This tendency can be seen in all Airsheds, though is less pronounced for WBEA stations for the 1-R metric than for the other Airshed stations. Although the averages result in different clusters compared to time-filtered data, the inclusion of short time spikes within a long-term average also results in a reduced abililty to identify locally unique source signals: the level of “smoothing” associated with averaging is sufficiently high that similarities with respect adjacent stations may be lost.

A consequence of this analysis is that data which consists of monthly averages (such as the passive data) will lack sufficient information to distinguish the extent to which station records within an Airshed are unique to that Airshed, using hierarchical clustering. On one hand, this shows a limitation of the analysis methodology – the uncertainty with regards to assigning relative levels of similarity and uncertainty becomes higher as the data are subjected to (or the result of) longer duration averaging periods. On the other hand, the analysis also shows an inherent issue with observations which are long-term averages: if one of the intentions in monitoring is to provide information on the relative influence of local sources of emitted pollutants versus regional background concentrations due to long-range transport, the use of long-term average observations (or long-term averages of short-term observations) loses that information through the process of including short-term events into a long-term average. For example, a single one-hour “spike” in SO_2 concentration which is 720 times greater than the typical background concentration, with background concentrations for the remainder of the month, is indistinguishable from a month long record with a continuous elevation above background of 0.14% ($1+720^{-1}$), when both are averaged to a monthly level.

This may in turn explain some of our findings with the bimonthly data for NO_2 and SO_2 in which stations located in different airsheds form clusters – the monthly averaging time of the underlying passive data is sufficiently long that the short term events which would separate out the Airsheds has been lost.

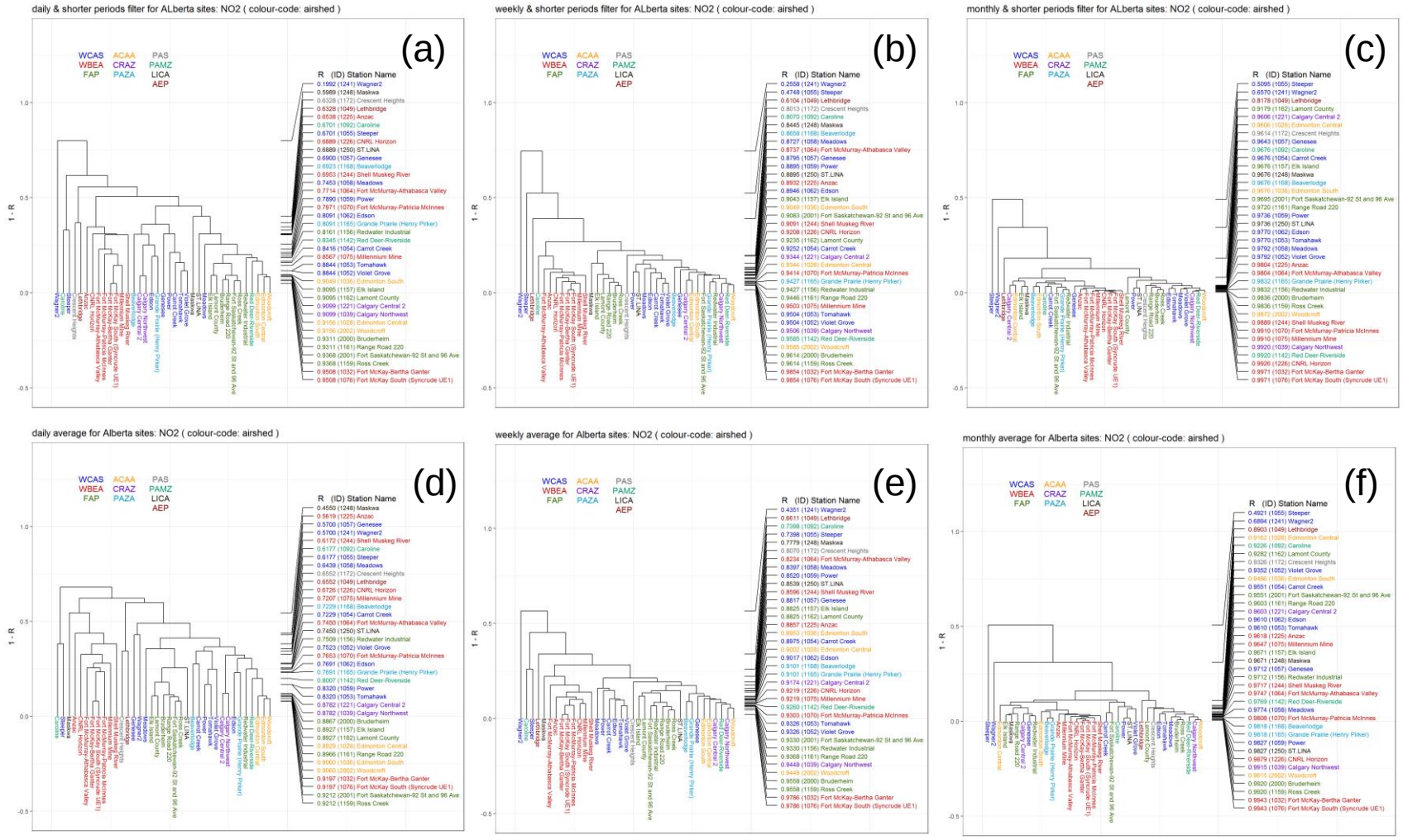


Figure 4.7 Continuous NO_2 1-R dendrogram analysis, averaging versus time-filtering. Top row: time-filtering, with (a) daily, (b) weekly and (c) monthly scales removed. Bottom row: time-averaging, with d) daily e) weekly f) monthly averages. Airshed names: WCAS: West Central Airshed Society, WBEA: Wood Buffalo Environmental Association, FAP: Fort Air Partnership, ACAA: Alberta Capital Airshed Alliance, CRAZ: Calgary Regional Airshed Zone, PAZA: Peace Airshed Zone Association, PAS: Palliser Airshed Society, PAMZ: Parkland Airshed Management Zone, LICA: Lakeland Industrial Community Association (LICA). Stations are colour-coded according to Airshed.

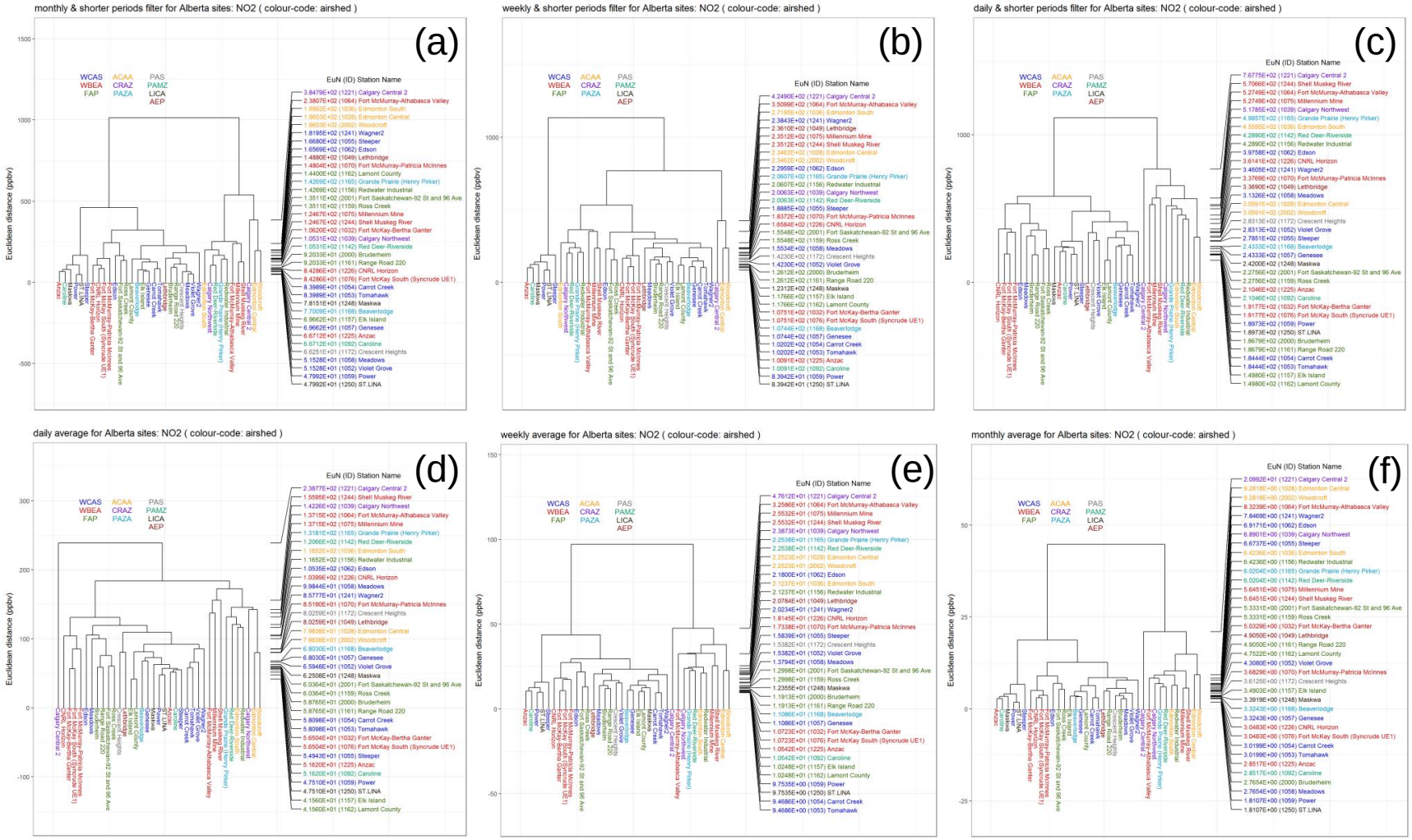


Figure 4.8 Continuous NO_2 Euclidean distance dendrogram analysis, averaging versus time-filtering. Panels arranged as in Figure 4.7.

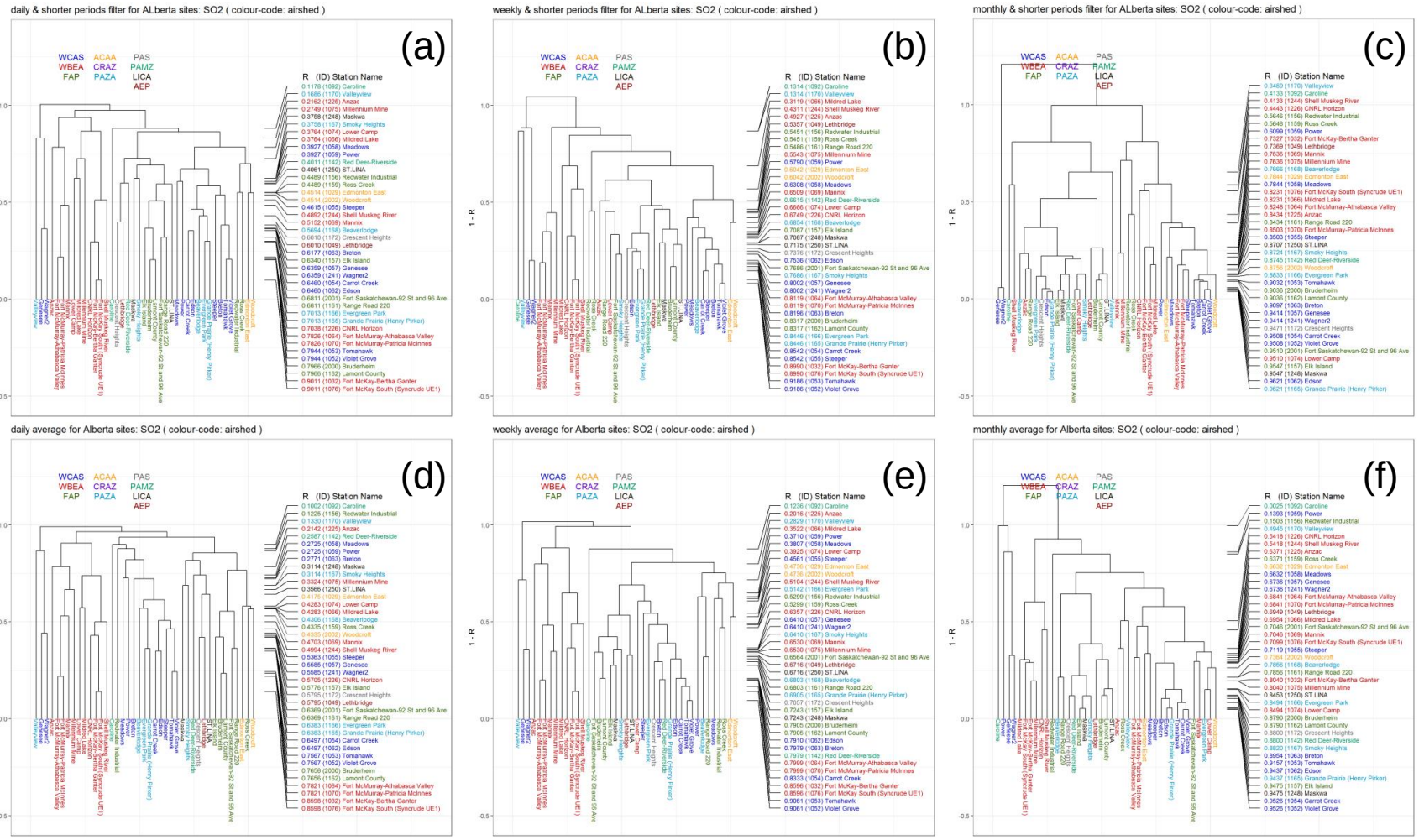


Figure 4.9 Continuous SO_2 1-R dendrogram analysis, averaging versus time-filtering. Panels arranged as in Figure 4.7.

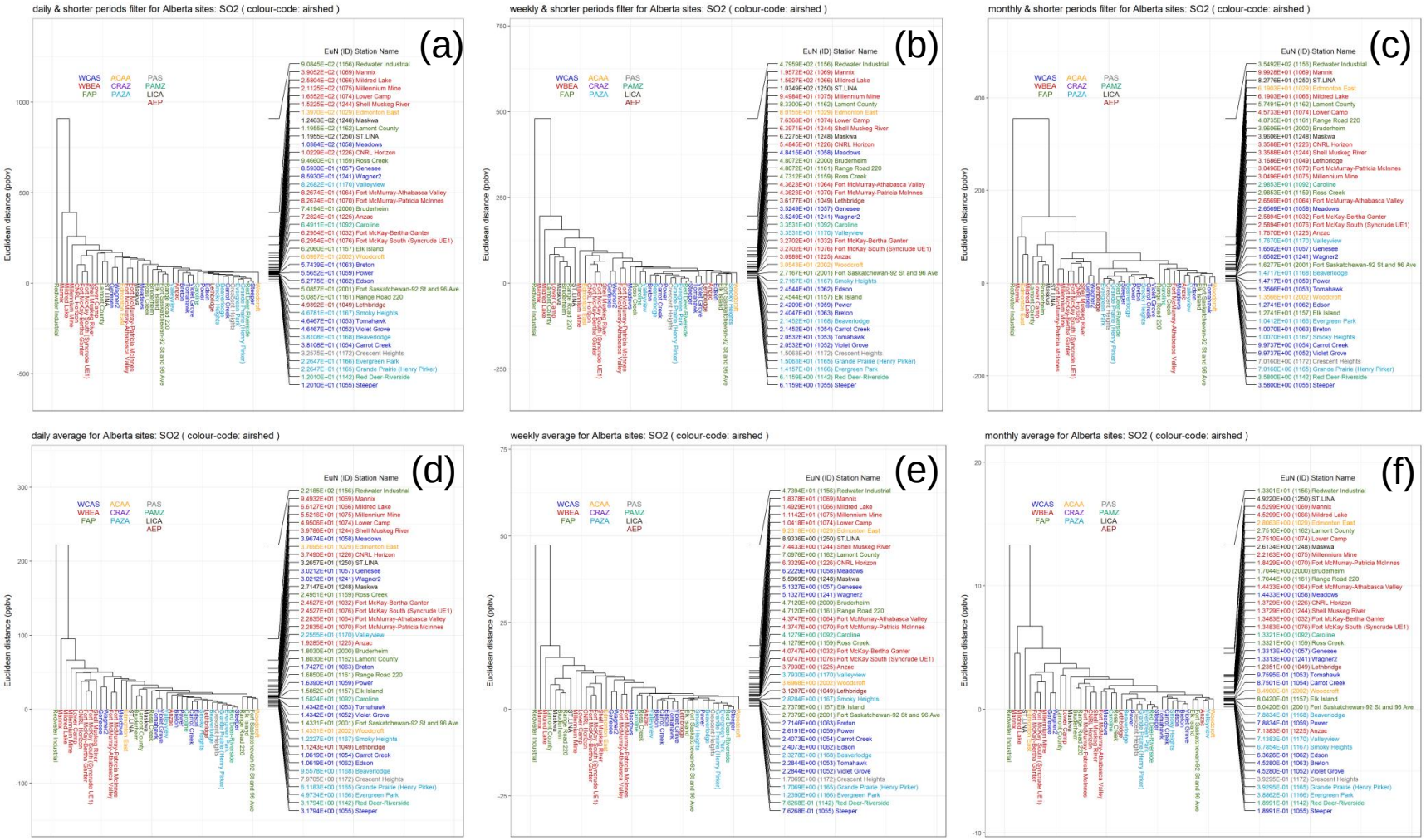


Figure 4.10 Continuous SO_2 Euclidean distance dendrogram analysis, averaging versus time-filtering. Panels arranged as in Figure 4.7.

4.3 The Effects of Random Error on Clustering

The passive and continuous data used for cluster analysis are subject to errors associated with the precision of the sampling methodology (see **Error! Reference source not found.**). We examine here the p otential errors associated with the detection limit of the monitoring methodology, using hourly time series at station locations from the GEM-MACH air pollution model (Makar *et al*, 2015, 2017, Moran *et al*, 2010). Model simulations from the period August 1, 2013 through July 31, 2014 were used to generate idealized "data" time series at observation station locations for three different chemical species, NO_2, O_3 and SO_2. Random noise was added to each of these time series, with the maximum magnitude of the noise for each species taken from the detection limit range of each instrument (1 ppbv for O_3 and SO_2, and 0.5 ppbv for NO_2). Dendrograms were then generated for each of the three chemical species, and each of the two sets of time series (with and without the random noise).

Comparisons of the resulting dendrograms for the original hourly time series, the corresponding time series with the added random noise, and the corresponding time series with monthly and smaller time scales removed, are shown in Figure 4.11 (O3), Figure 4.12 (NO2) and Figure 4.13 (SO2). The upper row of panels shows the dendrograms resulting from the original time series, and the lower row of panels shows the corresponding dendrogram for time series to which hourly random noise was added. The first two columns of panels in these figures correspond to the 1-R metric clustering for hourly and monthly and shorter time scales removed, respectively, and the 3rd and 4th column are the hourly and monthly and shorter time scales removed dendrograms for the Euclidean distance metric. Differences between the upper and lower rows panels in each column of the figures illustrates the extent to which the addition of random noise may affect the clustering – if differences can't be discerned, this impact is minimal. However, if the pattern of clustering changes between the upper and lower rows within a column changes, the impact of precision on the clustering results is larger.

The results show that for O_3, the addition of random error in the range +/- 1ppb has little impact on the clustering between stations, with the four dendograms in the lower panel remaining identical to the original results (Figure 4.11).

For NO2 (Figure 4.12), 1-R hourly data dendrograms for the original time series and the time series with the addition of random noise (a,e) seem identical, but this is not the case for the monthly and shorter time periods removed 1-R dendograms (b,f). For the Euclidean distance dendrograms, both the hourly and monthly and shorter time periods removed time series (c,d) of the original data appear identical to those from the analysis with the data bearing the addition of random noise ((g,h), respectively).

The SO_2 results (Figure 4.13) show the largest variation between the clusters generated with the original time series and those containing additional random noise. The difference in clustering is particularly noticeable for the 1-R dendograms, for both hourly (a,e) and time filtered data (b,f) and slightly less pronounced for Euclidean distances (c,g, and d,h, respectively).

This analysis suggests that the analysis methodology is least effected when the atmospheric concentrations measured are typically higher than the detection limit of the instrumentation carrying out the observations (the case for O_3, with +/- 1ppbv being relatively small compared to typical background ozone concentrations). However, for instruments measuring more discrete concentration events

(industrial plumes) interspersed with near-detection limit concentrations (e.g. the case for SO_2 with a detection limit of 1 ppbv, and to a lesser extent, NO_2, with a detection limit of 0.5 ppbv) the impact of precision on the clustering results may be stronger.

There are two important implications to this test. The first is that for species with similar concentration characteristics as SO_2 (many samples close to the detection limit) poor precision close for samples close to the detection limit will have a large impact on the analysis, leading to potentially erroneous results. The second is that the detection limit and the precision at those levels *matters* more for these species – that is, identifying the corresponding stations as being within a common airshed becomes more difficult due to low precision close to the detection limit.

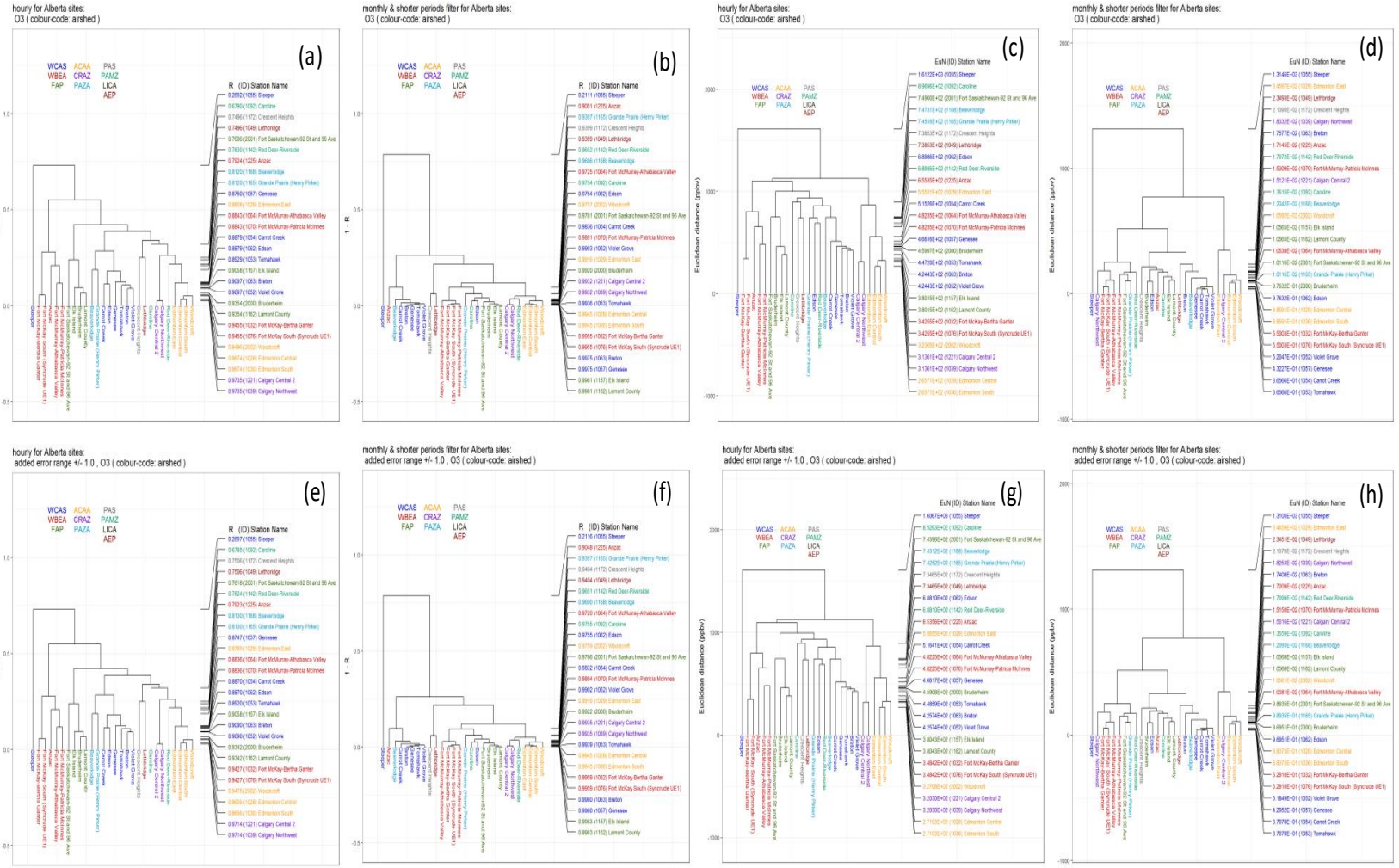


Figure 4.11 GEM-MACH O_3 concentrations predicted at the stations available for the 1 year study. Top row: dendrograms generated using the original model time series at each station; bottom row: random noise added at detection limit range of +/- 1 ppbv. 1-R metric results (a), (b), (e) and (f), Euclidean distance results (c), (d), (g), (h). Hourly data: (a), (e), (c), (g); Monthly and shorter time scales removed (b), (f), (d), (h). Pairs within each of the four columns of dendrograms can be compared – significantly different dendrograms within a given column indicate a greater impact of random nois

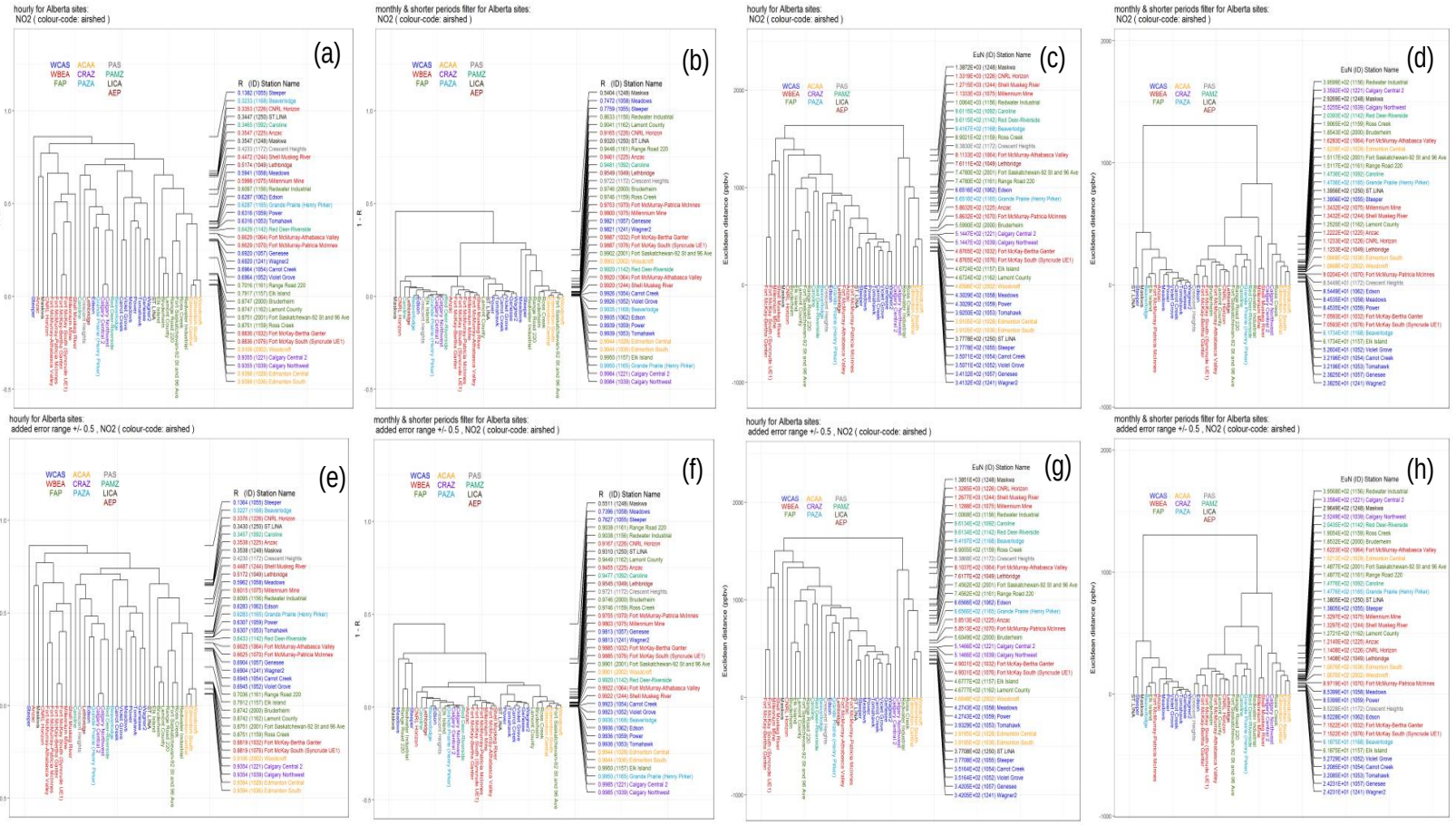


Figure 4.12 GEM-MACH NO_2 concentrations predicted at the stations available for the 1 year study. Rows and columns arranged as in Figure 4.11. Random noise level added +/- 0.5 ppbv.

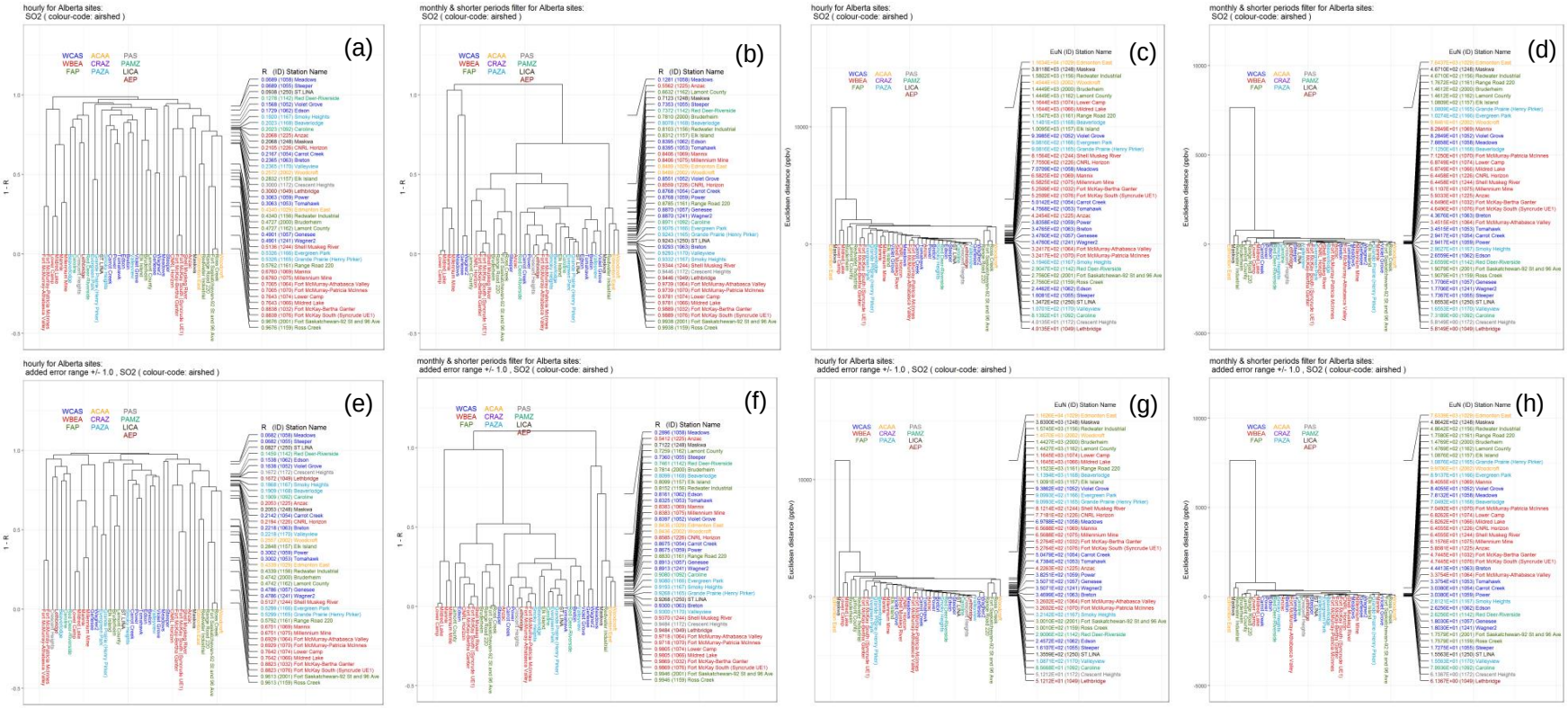


Figure 4.13 GEM-MACH SO_2 concentrations predicted at the stations available for the 1 year study. Panels arranged as in Figure 4.11. Random noise level added: +/- 1 ppbv

5 Summary

The methodology proposed in this report serves as a tool for providing information on the data collected at monitoring stations, and on their geographical location networks. The methodology expands on the work of Solazzo and Galmarini (2015), and includes two dissimilarity metrics for hierarchical clustering of monitoring data: the Euclidean distance and 1-R. The Euclidean distance metric allows cross-comparison of the stations in terms of the magnitude of the concentrations, whereas 1-R evaluates their temporal variation similarity. A KZ filter was adopted in its original low-pass configuration, filtering the original hourly time series to remove time scales periods less than daily, weekly and monthly, in order to distinguish the relative impacts of the different time scales on clustering.

The study suggests that optimization of networks should be carried out according to species rather than stations, as the species examined here are primarily emitted by different sources and/or the results of secondary chemistry. Overall, the methodology is able to identify groups of stations, which are influenced by common emissions sources (e.g. stations which are influence by oil sands emissions as opposed to stations located elsewhere) when the methodology is applied to hourly and, to some extent, daily time-filtered time series. Stations mainly influenced by seasonality are identified when the methodology is applied to weekly and monthly time-filtered data. The methodology also identifies monitoring stations making use of different monitoring methodologies (passive vs continuous) or monitoring stations records which are markedly different from all others in a given dataset due to outliers or data inaccuracy.

Stations will be similar when showing low 1-R (high correlation) and/or low Euclidean distance levels. Generally, we recommend evaluating the similarity of the station and their potential redundancy using both metrics when possible. However, both metrics can be used together or separately for the same purpose. The metric chosen for determining redundancies may thus be dependent on which of these two factors is considered to be more important with regards to the intent of the monitoring network, and combinations of the metrics may be preferred in assessing redundancies.

Ordering the station according to their similarity provides a relative ranking of similarity, depending on the available observation data (number of stations and chemical species observed) and time period analyzed, thus absolute thresholds for redundancy cannot be generated. In addition, other considerations such as spatial proximity to highly populated locations or sensitive ecosystems, the regulatory purpose of the station(s), and logistics (e.g. accessibility or power supply), may outweigh the recommendations based on similarity.

The methodology is highly dependent on the data available: number of stations and their spatial representation and the quality of the reported data. Therefore, we note that the lack of useable data may also be a potential consideration for network optimization. Regarding the quality of the data, in this analysis it should be mentioned the following confounding factors:

- Sampling accuracy. This seems to be related more to the passive monitors, with collocated passive monitors sometimes having large differences in Euclidean distance and/or large values of the 1-R

dissimilarity metric, when zero would be the expected value for collocated instruments sampling the same air.

- Averaging time of observations. Analyses in which hourly data were time-filtered to remove successively longer timescales suggested much of the information identifying pollutants sampled within specific Airsheds as unique resides in shorter timescales. Analyses of averaging time showed that increasing averaging times increases the degree of similarity, but at the expense of being able to resolve station records as unique to a given location.
- Random errors within the data records. Numerical tests show how random errors in the observations can potentially change the associativity analysis results.

There are also constraints on the interpretation of the analysis which must be mentioned:

- The analysis groups stations according to the degree of similarity but does not in and of itself provide the cause for that degree of similarity. The latter may only be achieved by examination of the data records, and the use of local knowledge of sources and conditions. Similarly, other constraints such as the availability of power and accessibility of station locations, and the intended purpose of the stations, while outside of the scope of the analysis carried out here, would also be constraints in network station siting.
- Passive and continuous monitors were analyzed together in order to determine the relative comparability of the two methodologies, but that portion of the analysis is not intended for providing information on relative levels of similarity and hence redundancy for continuous monitors; the separate analysis on continuous monitors alone should be consulted for the latter purpose.
- We have shown that averaging time may have an impact on the clustering results and longer term averages may lose some information which would shorter time scale averages would include - the methodology has the maximum benefit in assessing redundancies when the maximum amount of information is available (hourly data).
- The analysis is limited to the available stations which meet the data completeness criteria – some stations have been excluded due to data being insufficiently complete for analysis, and the analysis may be limited by the accuracy (precision) of the methodologies being used for data collection.

For each application of the methodology presented here, we provide caveats on the accuracy of the observation data, and recommendations on how the data may be used as an aid in assessing station redundancy:

WBEA Passive + Continuous NO_2 monitors

a) Caveats on the accuracy of the observation data:
   - Passive and continuous monitors separate out using the 1-R dissimilarity metric, and collocated continuous and passive monitors fail to cluster with each other, indicating that the two types of sampling have sufficiently different results that they are poorly comparable.
   - Passive monitors in general have smaller values of the Euclidean distance with each other than continuous monitors have with each other. This may represent an inability on the part of the passive monitors to accurately capture the magnitude of short-term events.
   - We note that the some of the passives in the WBEA Airshed are located above the local vegetation height, and this may affect the clustering with other monitors.

b) Using the available data
   - Table 4.1: The stations are ranked by 1-R (or R) and Euclidean distance – for NO2, the order of these rankings is almost reversed between the two metrics: stations which are the most redundant from the standpoint of correlation may be less redundant from the standpoint of Euclidean distance.

The choice of which metric to use in assessing potential relative redundancy thus depends which metric most closely represents the intent of the monitoring observations. Table 4.2 and Table 4.3 further illustrate this issue.

- The rankings of Table 4.1 could be used to determine potential redundancy through (a) choosing the station(s) at the bottom of the table (highest R or lowest Euclidean distance, depending on the metric considered most relevant), then identify that station on the relevant dendrogram, in order to determine the stations with which it clusters the most closely. One may thus see which stations would remain, representing the given station to that level of the metric of similarity, if it was removed from the list of stations (or moved to another location where its relative level of redundancy might be lower).

WBEA Passive + Continuous SO2 monitors

a) Caveats on the accuracy of the available data:
   - There was a lower tendency for the stations to cluster according to passive versus continuous technology compared to NO2. However, the clustering pattern did not always follow Airshed locations spatially for the 1-R metric, suggesting there may be a high degree of error in some of the observations (Figure 3.4) the Euclidean distance (Figure 3.5) sometimes showed a “close to site” versus “far from site” clustering for high EuN values (Figure 3.5 (b)).
   - 1-R and Euclidean distance metrics had a greater tendency to agree on redundancy levels than was the case for NO2, i.e. frequently the same stations had relatively high correlation coefficients and low Euclidean distances.
   - Note that the later use of model values as a surrogate for observations suggests that of the low accuracy of the SO2 sampler has a very strong impact on the clustering behavior – the low precision in sampling makes the data less useful, and harder to interpret.
b) Using the available data:
   - The rankings of Table 4.4 could be used to determine potential redundancy through a similar process as described for Table 4.1, above.

LICA Passive + Continuous NO2 monitors

a) Caveats on the data
   - The NO2 monitors had a more common ranking of similarity between both metrics than was seen for WBEA sites (perhaps fewer local sources/more distributed sources for LICA).
   - Collocated continuous and passive monitors had non-unity correlations (range 0.22 to 0.44) and sometimes quite large Euclidean distances (range 3.9 to 10.7 ppb). This indicates a large degree of incommensurability between the two measurement technologies. Collocated passive monitors had high Euclidean distances as well (3.2 ppb), indicating a high level of noise in the passive observations. Table 4.6 and Table 4.7 illustrate this issue.
b) Using the available data
   - Table 4.5 may be used to rank stations based on redundancies for the 1-R and Euclidean distances, similar to Table 4.1, above.

LICA Passive + Continuous SO2 monitors

a) Caveats on the data
   - Both Euclidean and 1-R rankings agreed in the general trend (similar stations appeared in the bottom of Table 4.8.
   - Continuous monitors tended to correlate better with each other than with (sometimes collocated) passive monitors.
   - Collocated passive monitors had non-zero Euclidean distances and 1-R values, though Euclidean distances were smaller than for WBEA stations, indicating a greater degree of redundancy for this metric in LICA than WBEA.
   - Table 4.9 and Table 4.10 show that several highly correlated station pairs also have relatively low Euclidean distances despite separations of up to 51 km.
   - The analysis using model data degraded due to adding noise to the data suggests that SO2 may be strongly impacted by sampling inaccuracy.

b) Using the available data
   - Table 4.8 may be used to rank stations in a similar manner to Table 4.1, above.

All Alberta Passive + Continuous NO2 monitors

a) Caveats on the Data
   - Stations that are widely separated in space may have similar time variation due to similar emissions sources nearby (e.g. mobile emissions of NOX being the dominant factor in NO2 1-R clustering, coal-fired powerplants with similar seasonal power loads and hence similar 1-R time series clustering for SO¬2).
   - Stations with the highest correlations are not necessarily the ones with the lowest Euclidean distances – 1-R is the metric most relevant for station time series “shape” while Euclidean norm is the most relevant station for concentration magnitudes.
   - Stations clustering with the lowest Euclidean distances are often sampling “background” air; e.g. mountain top sites and other remote locations.
   - NO2 is to some extent affected by adding random error to the concentrations – this makes it difficult to distinguish different sites, and low Euclidean norms and poor correlation coefficients may reflect inaccuracy of the sampling methodology.
   - Continuous stations remain distinct from passive stations to 1-R levels of 0.5 – this indicates that the methodologies are not really equivalent, but have systematic or random differences. Aside from the possible inaccuracy of the sampling methodology mentioned above, this seems to reflect inaccuracies for the passive stations, several of which are collocated with continuous stations, yet fail to correlate with them. Collocated passive monitors have Euclidean distances as high as 3.2 ppb, and collocated continuous and passive monitors have Euclidean distances as high as 14.4 ppb.

b) Using the available data,
   - Table 4.11 may be used to provide relative similarity rankings as an aid to assessing redundancy, in a similar manner to Table 4.1, as described above.

All Alberta Passive + Continuous SO2 monitors

a) Caveats on the Data
   - The 1-R and Euclidean distance metric rankings differ significantly – decisions must be based on the metric most aligned with the purpose of the monitoring network.
   - Loss of within-Airshed 1-R clustering tends to occur at higher correlation levels, potentially indicating a greater dependence on very local emissions sources.
   - Correlations are generally lower than for SO2, again likely the result of the nature of the emissions sources (large stacks).

b) Using the available data
   - For this dataset, an additional analysis was carried out which combined both metrics: The 1-R and Euclidean distances are ranked in 1-R clusters occurring within Airsheds were retrieved from the 1-R dendrogram and the corresponding maximum and minimum Euclidean distance between members of the cluster can be retrieved from the Euclidean dendrogram. This allowed groups of stations with relatively high correlations and low Euclidean distances to be identified (Table 4.13). WBEA 1-R clusters ranked in this fashion tended to have higher Euclidean distances (i.e. are less similar, less redundant), while six clusters (one in PAZA, PAS, FAP, LICA and two in PAMZ) had both higher 1-R values and Euclidean distances, indicating a greater degree of redundancy.
   - Single metric redundancies may be assigned based on the relative rankings of Table 4.12; the data may be used in assessing potential redundancies as described for Table 4.1, above.

All Alberta Continuous Monitors

a) Caveats on the Data
   - Analyses carried out in Section 4 suggest that the hourly data hold the most information for useful similarity rankings and should be the focus for redundancy assessment. For a given time scale, the magnitudes of both metrics vary widely between species – this finding shows that redundancies must be considered within each species separately; stations which may be more redundant (for either 1-R or Euclidean distance) for one chemical species may be much less redundant for a different chemical species.

b) Using the available data
   - The dendrograms in Figures 3.15 to 3.34 show 1-R and Euclidean distance rankings for the different stations, at the different timescales, with the higher R values and lower Euclidean distances identifying the more redundant stations within a given metric. Table 4.14 and Table 4.15 identify the highest and lowest ranking members of each Figure for the hourly data, and the rankings for a given time scale across all stations examined are provided to the right of each Figure.
   - The relative rankings appearing to the right of each of the hourly analyses in each of Figures 3.15 through 3.34 may thus be used to aid in assessing potential redundancies – these may be examined in the same manner as described for Table 4.1, above.

In this report, we described the time-filtering and cluster methodology chosen for the network optimization project (Phase 1), and its applicability for the optimization of the current monitoring network in Alberta (Phase 2). In Phase 3 of the network analysis project, the same methodology is applied to hourly model results extracted at station locations, to assess the model's ability to create matching associations between station records. In phase 4, the methodology is applied to gridded model output time series, treating each grid-cell as a potential monitoring station location, to generate maps describing dissimilarity sub-regions, within which a single station will represent the entire sub-region, to a given level of dissimilarity. These maps may be combined with other georeferenced data to assist in monitoring network design.

6 References

Alberta Environment and Sustainable Resource Development (AESRD). 2014. Development of Performance Specifications for Continuous Ambient Air Monitoring Analyzers. Government of Alberta, Alberta, Canada.

Alberta Environment and Parks (AEP). 2016. Air Monitoring Directive Chapter 4: Monitoring Requirements and Equipment Technical Specifications. Government of Alberta, AEP, Air, No. 1-4, Alberta, Canada.

ARC, Alberta Research Council (1998) Independent Validation of Chemex (Maxxam Analytics Inc.) All Season Passive Sampling System (CSPSS). Agreement #JPD 003.0097, March.

Bari, M.A.; Curran, R.L.T.; Kindzierski, W.B. 2015. Field performance evaluation of Maxxam passive samplers for regional monitoring of ambient SO_2, NO_2 and O3 concentrations in Alberta, Canada. Atmos. Environ. 2015, *114*, 39–47.

Brassard, R. (2001) Field validation of passive sampling devices and design of passive sampling networks. Proceedings, CPANS Voluntary Environmental Management Programs Conference", May 10–11, Edmonton, AB, Canada.

Bytnerowicz, A., Fraczek, W., Schilling, S. and Alexander, D. (2010). Spatial and Temporal Distribution of Ambient Nitric Acid and Ammonia in the Athabasca Oil Sands Region, Alberta. J. Limnol. 69: 11–21.

Cox, R.M. (2003). The Use of Passive Sampling to Monitor Forest Exposure to O_3, NO_2 and SO_2: a Review and Some Case Studies. Environ. Pollut. 126: 301–311.

Eskridge, R.E., Ku, J.Y., Rao, S.T., Porter, P.S., Zurbenko, I.G.: Separating different scales of motion in time series of meteorological variables. Bull Am Meteorol Soc 78, 1473–1483, doi:10.1175/1520-0477(1997)078<1473:SDSOMI>2.0.CO;2, 1997.

EPCM (2000): EPCM Associates Limited, 2000. Evaluation of Passive Sampling Systems at TEEM Jack Pine Monitoring Sites: Part II. Dry Deposition of SO_2. A Report to the Terrestrial Environmental Effects Monitoring Program of the Wood Buffalo Environmental Association, Calgary, AB.

Fraczek, W., Bytnerowicz, A., Legge, A., 2009. Optimizing a Monitoring Network for Assessing Ambient Air Quality in the Athabasca Oil Sands Region of Alberta, Canada. Alpine Space e Man & Environment. In: Global Change and Sustainable Development in Mountain Regions, 48, 7: 127-142.

Gerboles, M., Buzica, D., Amantini, L., Lagler, F., & Hafkenscheid, T. 2006. Feasability study of preperation and certification of reference materials for nitrogen dioxide and sulfur dioxide in diffusive samplers. *Journal of Environmental Monitoring* , 8: 174-182.

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd Edn. Springer, New York (2009).

Hogrefe, C., Rao, S.T., Zurbenko, I.G., Porter, P.S.: Interpreting information in time series of ozone observations and model predictions relevant to regulatory policies in the eastern United States. Bulletin of the American Meteorological Society 81, 2083–2106, 2000.

Hogrefe, C., Vempaty, S. , Rao, S.T., Porter, P.T.: A comparison of four techniques for separating different time scales in atmospheric variables. Atmos. Environ., 37 , 3, 313-325, doi: 10.1016/S1352-2310(02)00897-X, 2003.

Hsu, Y.-M., Percy, K., Hansen, M. 2010. Comparison of passive and continuous measurements of O3, SO2 and NO2 in the Athabasca Oil Sands Region. In: Proceedings of the 2010 (103rd) A&WMA Annual Conference. Air & Waste Management Association, Pittsburgh, PA.

Johnson R.A. and Wichern D.W. 2007. Applied Multivariate Statistical Analysis. Pearson Prentice Hall, Pearson Education Inc. Upper Saddle River, NJ, USA

Krupa, S.V. and Legge, A.H. 2000. Passive Sampling of Ambient, Gaseous Air Pollutants: an Assessment from an Ecological Perspective. Environ. Pollut. 107: 31–45.

Kirby, C., Fox, M., Waterhouse, J., Drye, T., 2001. Influence of environmental parameters on the accuracy of nitrogen dioxide passive diffusion tubes for ambient measurement. J. Environ. Monit. 3: 150-158.

Makar, P.A., Staebler, R.M., Akingunola, A., Zhang, J., McLinden, C., Kharol, S.K., Pabla, B., Cheung, P., and Zheng, Q., (2017) The effects of forest canopy shading and turbulence on boundary layer ozone, Nature Communications, 8:15243 | DOI: 10.1038/ncomms15243.

Makar, P.A., Gong, W., Milbrandt, J., Hogrefe, C., Zhang, Y., Curci, G., Zabkar, R., Im, U., Balzarini, A., Baro, R., Bianconi, R., Cheung, P., Forkel, R., Gravel, S., Hirtl, M., Honzak, L., Hou, A., Jimenez-Guerrero, P., Langer, M., Moran, M.D., Pabla, B., Perez, J.L., Pirovano, G., San Jose, R., Tuccella, P., Werhahn, J., Zhang, J., and Galmarini, S. (2015a). Feedbacks between air pollution and weather, Part 1: Effects on weather, Atm. Env., 115, 442-469.

Makar, P.A., Gong, W., Hogrefe, C., Zhang, Y., Curci, G., Zabkar, R., Milbrandt, J., Im, U., Balzarini, A., Baro, R., Bianconi, R., Cheung, P., Forkel, R., Gravel, S., Hirtl, M., Honzak, L., Hou, A., Jimenez-Guerrero, P., Langer, M., Moran, M.D., Pabla, B., Perez, J.L., Pirovano, G., San Jose, R., Tucella, P., Werhahn, J., Zhang, J., and Galmarini, S. (2015b). Feedbacks between air pollution and weather, part 2: Effects on weather, Atm. Env., 115, 499-526.

Moran, M.D., Menard, S., Talbot, D., Huang, P., Makar, P.A., Gong, W., Landry, H.,Gravel, S., Gong, S., Crevier, L.-P., Kallaur, A., Sassi, M. (2010). Particulate-matter forecasting with GEM-MACH15, a new Canadian air-quality forecast model. In: Steyn, D.G., Rao, S.T. (Eds.), Air Pollution Modelling and its Application XX, Springer, Dordrecht, pp. 2890-292

Næs T., Brockhoff, P.B., and Tomic, O. (2010). Statistics for Sensory and Consumer Science, 6th edition, John Wiley & Sons, Ltd, Wiltshire,UK. ISBN: 9780470518212

Palliser Airshed Society (PAS). 2016. A Year in the Palliser Airshed – 2006 Annual Report. Medicine Hat, Alberta, Canada.

Partyka, M., Zabiegala, B., Namiesnik, J., Przyjazny, A., 2007. Application of passive samplers in monitoring of organic constituents of air. Crit. Rev. Anal. Chem. 37: 51-78.

Pippus, G.J. 2012. Assessment of Sources of Uncertainty in Passive Samplers of Ambient Air Quality: Evaluation Lakeland Industry and Community Association Airshed 2009-2011. M.Sc. thesis report. Royal Roads University, Victoria, BC

Salem, A., Soliman, A., El-Haty, I., 2009. Determination of nitrogen dioxide, sulfur dioxide, ozone, and ammonia in ambient air using the passive sampling method associated with ion chromatographic and potentiometric analysis. Air Qual. Atmos. Health 2: 133-145.

Seethapathy, S., Górecki, T. and Li, X. (2008). Passive Sampling in Environmental Analysis. J. Chromatogr. A 1184: 234–253.

Solazzo, E. and Galmarini, S.: Comparing apples with apples: Using spatially distributed time series of monitoring data for model evaluation, Atmos. Environ., 112, 234–245, 2015.

Tang, H. 1998. Development of all-season passive sampling systems — a summary. Research Report for Research Canada Council", Contract No. 027240U-PH8.

Tang, H. 2001. Introduction to Maxxam all-season passive sampling system and principles of proper use of passive samplers in the filed study. In Proceedings of the International Symposium on Passive Sampling of Gaseous Air Pollutants in Ecological Effects Research. *TheScientificWorld* 1: 463–474.

Tang, H., Brassard, B., Brassard, R, and Peake, E. 1997. A new passive sampling system for monitoring SO2 in the atmosphere. FACT 1(5), 307–315.

Tang, H., Lau, T., Brassard, B., and Cool, W. 1999. A new all-season passive sampling system for monitoring NO2 in air. FACT 6, 338–345.

WBK and Associates Inc (WBK). 2007. Field Precision and Accuracy of Maxxam Passive Samplers for NO2, O3, and SO2 Used in the Wabamun-Genesee Area Ambient Air Monitoring Program, p. 13. Final Report, St. Albert, AB.

Xiaoliang Wang, Judith C. Chow, Steven D. Kohl, Kevin E. Percy, Allan H. Legge and John G. Watson (2015) Characterization of PM2.5 and PM10 fugitive dust source profiles in the Athabasca Oil Sands Region, Journal of the Air & Waste Management Association, 65:12, 1421-1433, DOI: 10.1080/10962247.2015.1100693.

Yan, S. and Wu G. 2016. Network Analysis of Fine Particulate Matter (PM2.5) Emissions in China. Sci. Rep. 6, 33227; doi: 10.1038/srep33227

Zabiegala, B., Kot-Wasik, A., Urbanowicz, M., Namiesnik, J., 2010. Passive sampling as a tool for obtaining reliable analytical information in environmental quality monitoring. Anal. Bioanal. Chem. 396, 273-296.

Appendix

B. The KZ Filter, Low-Pass versus Band Pass Filtering

The KZ-filter is defined as an iteration of a moving average filter applied on a time-series $S(t_i)$ (Zurbenko, 1986):

$$KZ_{m,p} = R_{i=1}^{p}\left\{J_{k=1}^{W_i}\left[\frac{1}{m}\sum_{j=\frac{m-1}{2}}^{\frac{m-1}{2}} S(t_i)_{k,j}\right]\right\}\{\, W_i = L_i - m + 1 \qquad \text{(A1)}$$

Where R is the iteration, m is the window size, p is the number of iterations, J is the running window, $S(t_i)$ the time series, and L_i is the length of the time series $S(t_i)$. Equation (A1) may be interpreted as p successive applications of a moving average of length m to the time series S, with the updated S being used as the starting time series for the subsequent moving average. The initial time series must thus have additional entries before and after the period L of interest, in order to result in a filtered time series of length L following the last application of the moving average. The first moving average is computed with a running window J and becomes the input for the second pass, and so on. The KZ filter controlling parameters m and p allow different time scales to be removed and filtered, as is described below.

The KZ belongs to the class of low-pass filters (since it filters periods smaller than a selected cut-off represented by a specific pair of m and p. The filter removes high frequency variations from the data (with respect to the window size) and belongs to the class of low-pass filters (since it filters periods smaller than the selected cut-off period). The KZ filter's original intent was a low-frequency pass filter but has been used as a band-pass in several air quality applications (e.g. Kang *et al* (2013), Galmarini *et al* (2013), Hogrefe *et al* (2000), Rao *et al*, (1997)), through taking the differences of time series pre-filtered for different time scales. However, the application of the difference in KZ filters for band-pass purposes does not separate the spectral components completely, with the energy spectrum overlapping on between the neighbour components (Hogrefe *et al*., 2000, 2003). The band-pass applications of the KZ filter suggested by Solazzo and Galamarini (2015) were tested by the authors of the current report. Artificial time series were constructed to examine the band-pass application's ability to separate known time-scales in those time series; the results were mixed, with intermediate time scales known to be in the input data failing to be resolved in subsequent clustering analysis. The band-pass approach's inability to completely separate adjacent time scales is the likely cause of this problem; too much energy leakage occurred, reducing correlations in clustering, and adding "noise" to the analysis.

However, the KZ filter in its original low-pass form was found to be able to separate the time scales in the test data accurately, with clustering showing the influence of the different time scales, given an appropriate choice of the filtering parameters m and p. That is, the analysis used here removes all of the energy below each of the time scales of interest (or, equivalently, above specific frequency thresholds), rather than attempting a band-pass approach. In addition to the hourly QA/QC and gap-filled data, the KZ filter was thus used to remove the energy for periods less than 1 day ($KZ_{17,3}$), less than 7 days ($KZ_{95,5}$) and less than 30 days is removed ($KZ_{523,3}$).

The choice of the values of *m* and *p* for these filters follows from the energy characteristics of the filter system. These can be derived from the transfer function of the KZ filter (see Eskridge *et al*. (1997) and Zurbenko, (1986) for details on the transfer function), given by

$$\left|\emptyset_{m,p}(\omega)\right|^2 = \left[\frac{1}{m}\frac{\sin(\pi m\omega)}{\sin(\pi\omega)}\right]^{2p} \qquad \text{(A2)}$$

where ω has units of cycles per hour (frequency), for hourly observation data. The transfer function defines the energy passed or removed by the filter as a function of frequency. Figure A1 shows the lines defining the low-pass filters for (m,p) = (17,3), (95,5), and (523,3) used in the current analysis. The frequencies to the left of the lines are "passed" by the filter for the given value of (m,p), those to the right are removed.

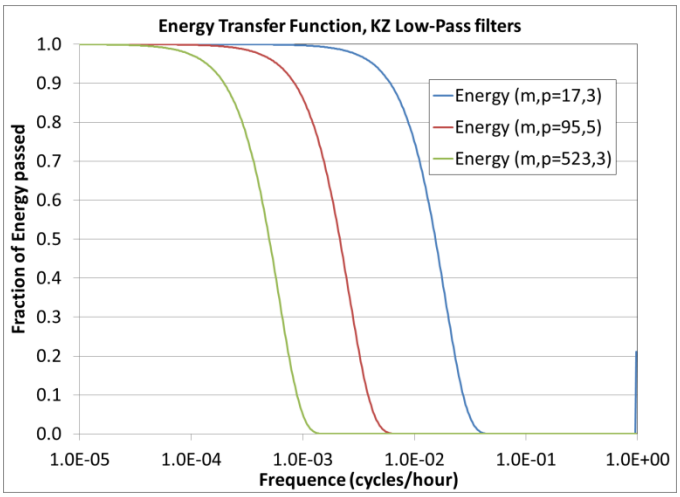


Figure A1 Energy transfer functions for KZ17,3, KZ95,5, KZ523,3

It can be seen from inspection of Figure A1 that the lines forming the boundaries between frequencies which are passed and those which are removed are not step-functions, but have a gradual change – for example, the (523,3) KZ filter passes 99.75% of the energy for frequencies less than $3.0\text{x}10^{-5}$ cycles/hour (period greater than 3.75 years), 50% of the energy at $5.01\text{x}10^{-4}$ cycles per hour (periods of 83.2 days) and less than 0.25% of the energy at frequencies greater than $1.36\text{x}10^{-3}$ cycles/hour (periods less than 31.1 days). The filter characteristics of the three low-pass filters used here are given in Table A1 below.

Table A1 shows that the three different (m,p) pairs selected for our work remove 99.75% of the energy for periods less than 1 day (17,3), 7 days (95,5), and 31 days (523,3), respectively. In subsequent figures and drawings, time series subjected to these filters will be referred to as having removed periods less than 1 day, less than 1 week and less than one month. It should also be noted that despite the gradual slope of each band-pass filter, the near-complete removal of energy for periods less than those in the final column of Table A1 is a well-defined quantity. One can say with good confidence then that the resulting filtered time series will have less than 0.25% of the energy remaining for periods less than the limits shown in the table. One caveat on that is the 17,3 "daily" filter, which shows some energy leakage

at periods specifically of the original time series (one cycle per hour); this daily filter will contain about 20% of the energy of periods equal to the original hourly time series interval.

Table A1 Frequency and period pass characteristics of the three KZ filters used here.

M	P	Frequency 99.75%	Period 99.75%	Frequency 50%	Period 50%	Frequency 0.25%	Period 0.25%
17	3	9.38332×10^{-4}	44.4 days	1.54338×10^{-2}	2.70 days	4.12341×10^{-2}	1.01 days
95	5	1.26846×10^{-4}	328 days	2.14594×10^{-3}	19.4 days	6.06578×10^{-3}	6.97 days
523	3	3.04471×10^{-5}	3.75 years	5.00840×10^{-4}	83.2 days	1.35751×10^{-3}	31.1 days

The gradual slope, rather than a square-wave cut-off, for the KZ *low-pass* filter, highlights a potential difficulty with the use of the past use of that filter for *band-pass* purposes (e.g. Hogrefe *et al*, 2000, Solazzo and Galmarini, 2015). The use of the KZ filter as a band-pass filter involves two steps. In the first step, the KZ filter is applied on the original data for two different sets of (m,p) pairs, resulting in two different filtered time series. In the second step, the difference between these time series at each time point is constructed (lower order pair filtered time series – higher order pair filtered time series, at each time step). The KZ filters used in these past applications are shown in Table A1 $KZ_{3,3}$, $KZ_{13,5}$, $KZ_{103,5}$, and $KZ_{310,7}$. The time series resulting from the difference between the original time series and $KZ_{3,3}$ is referred to as “intra-day” (periods less than 12 hours), whereas “diurnal” (periods between 12 hours and 2.5 days), “synoptic” (periods between 2 days and 21 days) and “long-term” (periods between 21 and 90 days) time series are formed from the differences $KZ_{3,3}$-$KZ_{13,5}$, $KZ_{13,5}$-$KZ_{103,5}$, and $KZ_{103,5}$-$KZ_{310,7}$, respectively. $KZ_{310,7}$ is said to form the “seasonal” (periods over 90 days) component of the time series.

The energy transfer functions for these “standard” filters are applied two ways in **Error! Reference s ource not found.**. **Error! Reference source not found.**(a) shows the low-pass filters for the regions bounded by $KZ_{3,3}$, $KZ_{13,5}$, $KZ_{103,5}$, and $KZ_{310,7}$. The energy response of the filters has a similar shape to those in **Error! Reference source not found.**, though the energy response of the $KZ_{3,3}$ filter can be seen t o have a more significant contribution near frequencies of 1 cycle per hour, suggesting a significant “leakage” of energy from short time scales with this filter. The regions used in previous work to describe different filter bands are labelled as noted above. While it can be seen by inspection of **Error! Reference s ource not found.**(a) that the energy associated with the difference between any two KZ filters will vary depending on frequency, the implications of that variation are more clearly displayed in **Error! Reference s ource not found.** (b), in which the differences between low-pass filters are used to define the band-pass filters used in previous work. **Error! Reference source not found.**(b) shows significant overlap in filtered e nergy between the “diurnal” (green), “synoptic” (purple), “long term” (light blue), and “seasonal” (red) filters. For example, the seasonal and synoptic filters both pass 47% of the energy at a frequency of 5.75×10^{-4} cycles/hour (72 days), and the seasonal and synoptic filters both pass 49% of the energy at 2.0×10^{-3} cycles/hour (21 days); the diurnal and synoptic filters both pass 49% of the energy at 1.58×10^{-2} cycles/hour (2.6 days) and the diurnal and intraday filters share the same boundary for frequencies greater than 4.17×10^{-2} (1 day), including the region near frequencies of 1 cycle/hour. Some energy

leakage occurs between the diurnal and synoptic filters as well, at the less than 0.05 level. The filters are not the ideal “square wave” associated with a band pass, but are subject to considerable overlap.

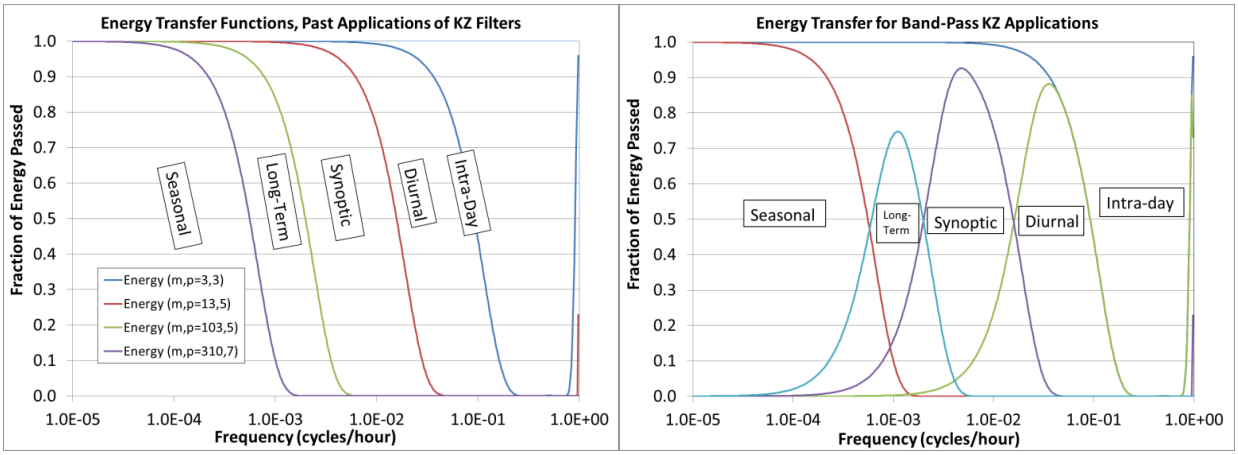


Figure A2 Energy transfer for previous applications of KZ filtering as a “band-pass” filter. (a) Transfer functions for low-pass filters. (b) Transfer functions for band-pass difference filters, as well as intra-day band-pass and seasonal low-pass filters

This degree of overlap has significant implications for the “bandpass” use of the KZ filter in the manner described in previous work (Eskridge *et al*, 1997, Hogrefe *et al*, 2000, Solazzo and Galmarini, 2015). The time labels for these filters are based on the 50% energy transfer levels of the differences to define a range in time represented by the filters. **Error! Reference source not found.**(b) shows that these b oundaries are not unique in energy – that is, a significant fraction of “seasonal” energy will be present in the “long-term” signal, a significant fraction of the “long-term” energy will be present in the “seasonal” signal, and so on.

In order to determine the potential impact of the overlap in band-pass on hierarchical clustering (described in more detail below), three time series were constructed for testing both low-pass and band-pass filtering combined with correlation analysis. The three time series are intended to represent hypothetical observations from three observation sites (A,B,C), and the time signals going into their construction are shown in **Error! Reference source not found.**, with the formulae describing the c omponents of each time series and the net time series shown in **Error! Reference source not found.**.

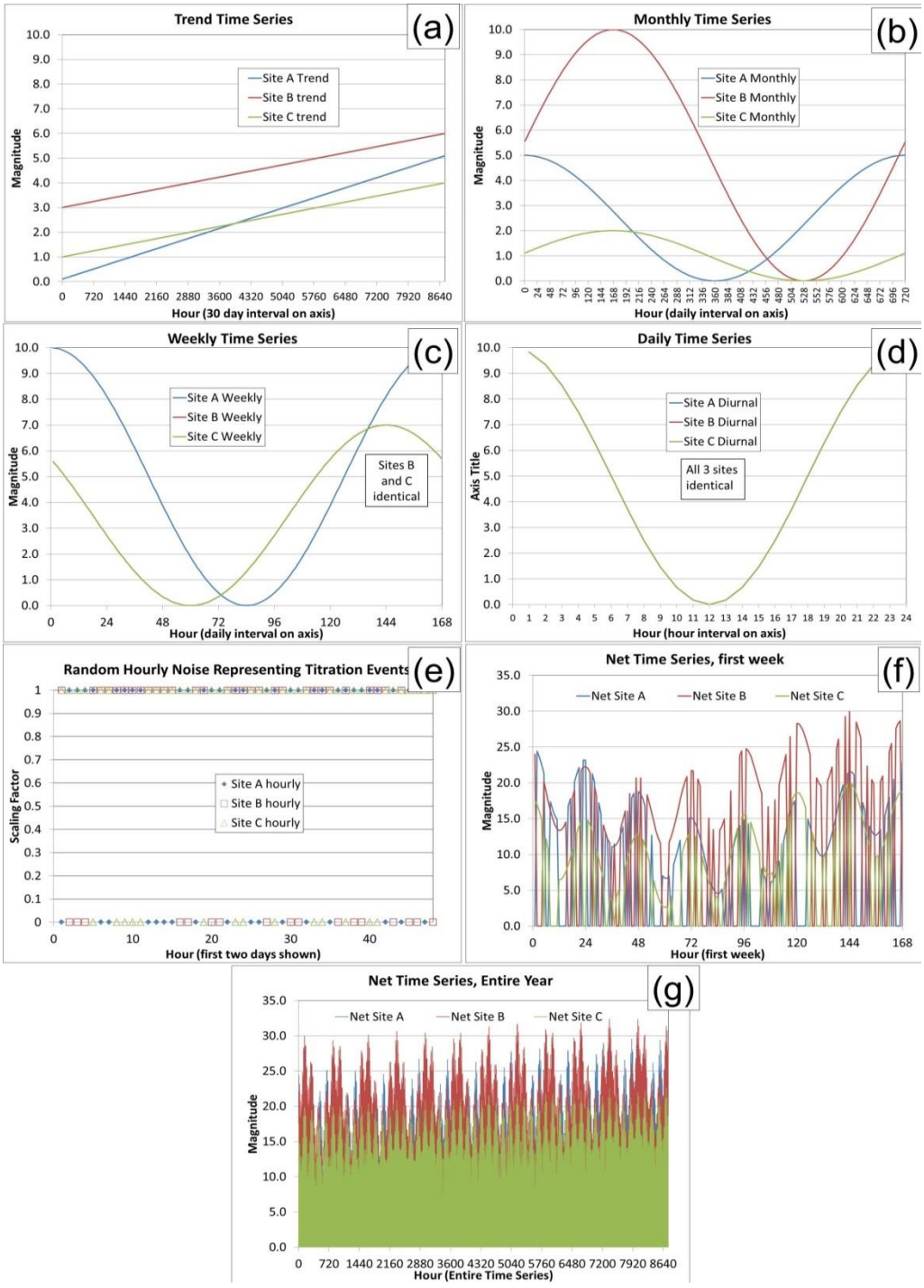


Figure A3 Construction of the test time series for three hypothetical stations. (a) Annual trend. (b) Monthly variation. (c) Weekly variation. (d) Diurnal variation. (e) Random noise (0 or 1) used to represent plume titration. (f) Resulting net signal. (g) Net signal for an entire year of hourly values.

Table A2 Components of Time Series for Testing

Site	Trend	Time Component Formula (h=hour of year) Monthly	Weekly	Daily
A	$h\left(\frac{5}{8760}\right)+0.1$	$2.5\left[cos\left(\frac{\pi h}{360}\right)+1\right]$	$5\left[cos\left(\frac{\pi h}{84}\right)+1\right]$	$5\left[cos\left(\frac{\pi h}{12}\right)+1\right]$
B	$h\left(\frac{3}{8760}\right)+3$	$5\left[cos\left(\frac{\pi(h-168)}{360}\right)+1\right]$	$3.5\left[cos\left(\frac{\pi(h+24)}{84}\right)+1\right]$	$5\left[cos\left(\frac{\pi h}{12}\right)+1\right]$
C	$h\left(\frac{3}{8760}\right)+1$	$\left[cos\left(\frac{\pi(h-168)}{360}\right)+1\right]$	$3.5\left[cos\left(\frac{\pi(h+24)}{84}\right)+1\right]$	$5\left[cos\left(\frac{\pi h}{12}\right)+1\right]$

In order to create the time series used for testing and displayed in **Error! Reference source not f ound.**(f,g), the four components for each site described in **Error! Reference source not found.** were added. The resulting time series where then multiplied by random numbers whereby at any given hour, two of the three time series values were multiplied by unity, with the remaining site value by zero, the choice of which time series to locally zero being chosen at random. This addition of random zeroing was added to mimic plumes which may reach only one station at a time (e.g. the summed time series representing ozone, and the zeroing representing a plume of NOx titrating ozone at one station and not the others). Error! Reference source not found. and **Error! Reference source not found.** show that all th ree stations are identical in terms of their diurnal variation, stations B and C are identical for the weekly variation, stations B and C have the same monthly variation and a magnitude offset, and the stations all have different long-term trends. Having constructed this test dataset of three stations, it may be used for different KZ filtering approaches in order to determine whether those approaches may discern the timescales known to exist within the constructed time series.

The aim of the dissimilarity analysis (described in more detail in the following section) is to compare station time series based on a metric such as 1-R, where R is the Pearson correlation coefficient, in order to group stations based on the lowest level of dissimilarity (or highest correlation). For a simple set of only 3 stations such as has been constructed here for testing, the correlation between their time series need only be calculated three times for the three unique pairs of the stations ((A,B), (B,C) and (A,C)). The different methodologies for KZ filtering are first applied to the time series, and then correlations are calculated for the three resulting pairs of filtered time series; these may be used to determine whether the methodology used recovers information about the time scale used. **Error! Reference source not found.** s hows this analysis using the 1-R metric, for the band-pass filters, starting from the time series constructed from **Error! Reference source not found.** and **Error! Reference source not found.**.

Table A3 1-R values between pairs of test time series, for original time series and KZ band-pass filters.

Original Hourly Time Series			**Interpretation**
	A	**B**	**C**	The dissimilarity (1-R) is greater than unity for all pairs – the addition of the random zeroing has created sufficient noise that the original time series are anticorrelated.
A	0.0000	1.298	1.260	
B	1.298	0.0000	1.233	
C	1.260	1.233	0.0000	
Intra-day dissimilarity (original time series - $KZ_{3,3}$)			**Interpretation**
	A	**B**	**C**	The intraday dissimilarity includes most of the random noise: since a different station is being zeroed at every hour, most of the noise appears in this time scale – the dissimilarity values are all greater than unity, indicating that most of the noise occurs at this time scale.
A	0.0000	1.489	1.503	
B	1.489	0.0000	1.466	
C	1.503	1.466	0.0000	
Diurnal dissimilarity ($KZ_{3,3}$ – $KZ_{13,5}$)			**Interpretation**
	A	**B**	**C**	The diurnal component is usually assumed to retrieve signals between 0.5 to 2.5 days. However, despite the identical diurnal signal present in all three original time series, the dissimilarity between all three time series remains high (the correlation remains low). The dissimilarity pairs ordered from lowest to highest are (A,C), (B,C), (A,B): the conclusion from this analysis would be that (A,C) are the most similar stations at this time scale, followed by (B,C) then (A,B). However, the temporal variation used to construct the time series is identical at this time scale – the band-pass methodology would lead to an erroneous conclusion. The low correlation is likely due to the lower frequency end of the band-pass including the time scale incorporating most of the noise.
A	0.0000	1.032	0.8968	
B	1.032	0.0000	0.9458	
C	0.8968	0.9458	0.0000	
Synoptic dissimilarity ($KZ_{13,5}$ – $KZ_{103,5}$)			**Interpretation**
	A	**B**	**C**	The synoptic component is usually assumed to retrieve signals between 2.5 and 21 days. The pairings here from lowest to highest dissimilarity are (B,C) < (C,A) < (A,B). The methodology has successfully identified the (B,C) pair as the most similar; from Table A1, this is correct – the weekly signal is identical for this pair. The other two pairs should be equally dissimilar based on Table A1, but this is only true to the first digit in the band-pass analysis.
A	0.0000	0.7909	0.7222	
B	0.7909	0.0000	0.5286	
C	0.7222	0.5286	0.0000	
Long-term dissimilarity			**Interpretation**

$(KZ_{103,5} - KZ_{310,7})$	
	A B C A 0.0000 0.8408 0.6003 B 0.8408 0.0000 0.4364 C 0.6003 0.4364 0.0000
	The long term dissimilarity is intended to isolate signals between 21 and 90 days. The Monthly signal from Table A1 should therefore be resolved by this analysis. From Table A1, the (B,C) pair should have the greatest degree of similarity – and this is reflected in the analysis. (A,C) is shown to have a greater degree of similarity than (A,B). The periodicity differences between (A,C) and (A,B) should be identical, but the average value of the signals (A: 2.5, B: 5.0, C: 1.0) are likely why (A,C) has been identified as being more similar than (A,B). So at this time scale the results are reasonable.

Error! Reference source not found. suggests that the diurnal filter may be severely affected by energy l eakage from other parts of the frequency spectrum, failing to identify the identical similarity in the diurnal signal constructed here (and indicating a low degree of similarity at that time scale in general). The synoptic dissimilarity correctly identified the most similar pair, but failed to give the remaining two pairs identical similarities, indicating that energy leakage from the adjacent bands are also present at this time scale. The long-term dissimilarity seems to have captured the main features of that signal; a combination of similarities in magnitude or phase lag of the monthly time series.

Error! Not a valid bookmark self-reference. provides the low-pass filter results for the filters described in **Error! Reference source not found.** and **Error! Reference source not found.**. The use of the KZ as a low-pass filter as in Table A4 has some advantages for the shorter time scales compared to the band-pass filters of Table A3 – the noise leakage from the short term variations has contaminated the band-pass filters for the diurnal signal, creating negative correlation coefficients, reducing correlations and obscuring the identical variation at that time scale. The synoptic band-bass similarity also shows some energy leakage. The low-pass filters have removed the high frequency noise due to the choice of *m* and *p* values. The interpretation between band-pass and low-pass filters of course differs – the low-pass includes all frequencies less than the cut-off frequency (or all periods greater than the cut-off period), and must be interpreted in that context. Here we choose to use the low pass filters for our subsequent analysis, largely to avoid the high frequency noise and energy overlap issues shown below.

Table A4 1-R values between pairs of test time series, for KZ low-pass filters.

Filters out time scales less than 1 day ($KZ_{17,3}$)				**Interpretation**
	A	**B**	**C**	The daily time (and shorter) variation has been removed. (B,C) are the most similar due to their identical weekly time series and time variation for the monthly time series. (A,B) are the least similar due to their difference in magnitude and period at both monthly and weekly time scales. (A,C) are intermediate due to the similar period at monthly time scales and the relatively small size of the offset at weekly time scales.
A	0.0000	0.8236	0.7208	
B	0.8236	0.0000	0.5282	
C	0.7208	0.5282	0.0000	
Filters out time scales less than one week ($K_{95,5}$)				**Interpretation**
	A	**B**	**C**	The highest similarity is between (B,C), suggesting that the trend and the magnitude of the monthly signal dominates the similarity. (A,B) are the least similar, indicating that the monthly period offset between the signals and the difference in slope in the trend between these stations results in lower similarity than between (A,C). The intermediate values of the latter reflect the identical periods of the monthly signal and the identical slope in the trend.
A	0.0000	0.8455	0.6172	
B	0.8455	0.0000	0.4377	
C	0.6172	0.4377	0.0000	
Filters out time scales less than 1 month ($KZ_{532,3}$)				**Interpretation**
	A	**B**	**C**	The dissimilarities are low (and hence the similarities are high) for all variable pairs. (A,C) are the most similar, reflecting the similarity in the magnitudes of these lines during the year. (B,C) are the next most similar pair, reflecting the similarity in the slopes. (A,B) are the least similar pair, reflecting the similarity in the slopes but the constant offset between this last pair.
A	0.0000	0.2855	0.1083	
B	0.2855	0.0000	0.2146	
C	0.1083	0.2146	0.0000	

A McLauren series expansion of the sinusoids in Equation (A2), to the first two terms in the expansion for the numerator and denominator functions may be used to approximate the frequency energy cut-off-curves. If A is the fractional energy passed at a given frequency in the line, then that frequency may be approximated by:

$$\omega_0 \approx \frac{\sqrt{6}}{\pi} \sqrt{\frac{1-(A)^{\left(\frac{1}{2p}\right)}}{m^2-(A)^{\left(\frac{1}{2p}\right)}}} \qquad \text{(A3)}$$

where ω_0 is the desired separating frequency and the approximate solution to the equation

$$\left|\emptyset_{m,p}(\omega)\right|^2 = A \quad \text{(A4)}$$

A value of A = ½ has been used in band-pass applications in the past to indicate the "boundaries" of these filters, though it can be seen from **Error! Reference source not found.**(b) and **Error! Reference source not found.** that the band-pass applications have significant energy leakage beyond these bounds.

The $KZ_{17,3}$, $KZ_{95,5}$, and $KZ_{523,3}$ filtering was then applied to the hourly data available as continuous observations. The continuous and passive bimonthly data was used *as is*, as the high frequency time scales have been naturally removed when averaging was applied. These KZ-filtered (for the continuous data) and unfiltered (bimonthly) time series were then analyzed using hierarchal clustering, described in more detail below.

B. Dissimilarity Analysis using Hierarchical Clustering: Mathematical Underpinning

Dissimilarity analysis comprises a group of methodologies used to rank datasets based on the extent to which they are different (or *dissimilar*) from each other. Here, the datasets are the time series of observations at different monitoring network stations. Highly dissimilar pairs of station datasets, or groups of datasets, are the least like each other, while station datasets with low levels of dissimilarity are the most like each other. Dissimilarity may thus be used to rank stations in terms of potential redundancy in that those stations having low levels of dissimilarity may be sufficiently similar to be redundant.

One of the most commonly used methodologies for dissimilarity analysis is hierarchical clustering; a well-established method to determine the inherent or natural groupings of datasets, and/or to provide a summarization of data into groups. The first step for hierarchal clustering is to choose a metric to describe how different (how dissimilar) a pair of time series are from each other. This metric is then calculated for all possible pairs of the time series comprising the dataset, resulting in a dissimilarity matrix. The matrix is then used to cluster the data based on their level of dissimilarity. The pair of time series with the lowest level of dissimilarity (i.e. are the most similar or closest to being identical according to the metric chosen) are combined in some fashion as a *cluster*. The metric of dissimilarity is then recalculated between this cluster and the remaining time series, the lowest dissimilarity pair is once again determined, resulting in another combination of time series and/or clusters calculated from previous iterations of the method. The number of clusters, which was originally equal to the number of time series in the original dataset, is thus reduced at each stage of the hierarchical clustering process, until only two clusters remain – once these are joined, the process is complete.

To recalculate the dissimilarity matrix based on the dissimilarity metric, here we make use of the *general averaging* method – once the initial level of dissimilarity between each of the time series comprising the original dataset have been calculated, and the lowest dissimilarity pair has been identified, the

dissimilarity between the new cluster and the remaining members of the dataset is represented by the *average of the metrics* between the two members of the data pair being brought together as the new cluster, with respect to each of the remaining members of the dataset. This is known as the *general average* or *linked average* or *average linkage* method (c.f. Næs *et al*, 2010). An alternative (and older) approach would be to average each value at each time within the two time series to create a new time series, then explicitly recalculate the dissimilarity metric with the remaining members of the dataset. However, general averaging has been shown in the past to provide robust and accurate clustering, with a substantial reduction in the processing time required to generate clusters. The processing time for methods in which the dissimilarity metric is explicitly re-calculated tend to scale as the third power of the number of time series in the initial dataset, while those making use of approximations such as general averaging scale as the second power of the number of stations. Approximations such as general averaging are thus the norm in modern applications of hierarchical clustering.

Here, the hierarchical clustering thus provides a ranking of stations based on their degree of dissimilarity – those stations which group into clusters early in the process (i.e. at low levels of dissimilarity), have time series which are more “like” each other in terms of the dissimilarity metric chosen. Those which are least like each other do not group into clusters until later in the process, at higher levels of dissimilarity. There are also many possible choices for the dissimilarity metric. In the analysis which follows, we determine dissimilarity separately with two metrics, 1-R and the Euclidean norm (described in detail below), starting by finding the pair of station time series with the lowest value of the metric (i.e. the most similar time series), and merge these two to form the first cluster. As data series and clusters merge, their combination as well as their level of dissimilarity at the point of merging is called a *node.* The consequent merging of other time series and clusters is repeated until all the clusters are combined, here using the average-linkage method. The analysis proceeds from the most similar station time series, building clusters between station time series and earlier clusters, until all of the station data have been merged into clusters. The order in which stations merge, as well as the dissimilarity level at which they merge (i.e. the nodes for the clustering) are tracked, and are used to generate explanatory diagrams of the clustering known as *dendrograms*. Dendrograms show the pattern of linkages between nodes as the analysis progressed, with vertical lines representing the level of dissimilarity between stations time series or between station time series and clusters, and horizontal lines showing which time series or clusters have become linked as nodes. Dendrograms thus resemble the root system of a tree, with the most similar stations forming the lowest level of the smallest roots, and the two least similar clusters being linked at the top of the diagram as the trunk of the tree. Dendrograms have M-1 nodes, where M is the number of stations in the original dataset; that is, there will be M-1 linkages (nodes) formed from a dataset starting with M stations. For very large numbers of stations, the dendrograms become difficult to interpret, but in the work carried out here the number of stations is sufficiently small that they are useful aids in interpreting the dissimilarity analysis results.

B.1 Dissimilarity Metric: 1-R

Solazzo and Gamarini (2015) chose as their dissimilarity metric 1-R, where R is the Pearson linear correlation coefficient, in their application of dissimilarity analysis using hierarchical clustering for European and North American ozone data. For two time series $X_I(t)$ and $X_J(t)$ ($=X_{I,t}$ and $X_{J,t}$) available for stations *I* and *J*, the Pearson's correlation coefficient is defined as follows:

$$R_{X_{I,t},X_{J,t}} = \frac{cov(X_{I,t},X_{J,t})}{\sigma_{X_{I,t}}\sigma_{X_{J,t}}} = \frac{\sum_{t=1}^{N} X_{I,t},X_{J,t} - N\overline{X_I}\,\overline{X_J}}{\left[\sum_{t=1}^{N}(X_{I,t}-\overline{X_I})^2\right]^{\frac{1}{2}}\left[\sum_{t=1}^{N}(X_{J,t}-\overline{X_J})^2\right]^{\frac{1}{2}}} \quad \text{(B1)}$$

where *cov* is the covariance, σ the standard deviation, and N is the number of entries in the time series for stations *I* and *J*. Here, the time series $X_I(t)$ may be the hourly observations, the time series after applying the KZ filtering on the hourly time series, or the bimonthly averaged observations, described elsewhere in this report.

The Pearson's correlation coefficient describes the level of similarity of the shape of the two time series, and has been used successfully for dissimilarity analyses of air pollution network data in the past (Solazzo and Galmarini, 2015). However, this metric fails to capture changes in the magnitude of concentrations between two time series. For example, a pair of time series in which the entries of one member of the pair are all 1/100 of those of the other member of the pair will have a correlation coefficient of unity, missing the impact of the difference in magnitude. For this reason, our analyses are repeated with a second dissimilarity metric, the Euclidean distance.

B.2 Dissimilarity Metric: Euclidean Distance

Additionally, a dissimilarity matrix was determined by computing the Euclidean distance for the time series, where the distance between time series $X_I(t)$ and $X_J(t)$ is:

$$d_{X_{I,t},X_{J,t}} = \sqrt{\sum_{t=1}^{N}(X_{I,t} - X_{J,t})^2} \quad \text{(B2)}$$

While the previous metric (1-R) is unitless, the Euclidean norm expresses dissimilarities in the units of the time series, and is the net magnitude of the differences between the two time series. Low values of the Euclidean distance thus represent pairs of time series or clusters which are closer to being identical in terms of magnitude, while high values of the distance represent time series pairs or clusters which have very different magnitudes.