Triangular Tessellation Documentation

Part II – Clustering on Boundaries

by

J.A. Galt

Genwest Systems Inc

P. O. Box 397

Edmonds, WA 98020

Genwest Technical Publication 15-001

ABSTRACT

Many trajectory models used to study environmental distributions are formulated to track Lagrangian particles embedded in Eulerian fields. Where the Eulerian fields are used to represent advective and diffusive processes. A particular modeling frame-work of this type is GNOME (General NOAA Operational Modeling Environment) which is used by NOAA's Office of Response and Restoration Emergency Response Division for scientific support during spill events. This note is a follow on to the analysis of GNOME output described in Galt,2011, which dealt primarily with the Lagrangian-to-Eulerian transformation and presentation of model output representing floating pollutants in non-convex, multiply connected domains. This work will focus on the Lagrangian-to-Eulerian transformation and presentation of model output representing the beached or stranded pollutants. The final section then considers the use of Eulerian probability density fields to define the information “entropy” of a trajectory model and provides a brief discussion on how this approach characterizes physical processes represented in the model.

The transformation presented here provides a robust method to represent quantity of pollutant to impact a shoreline without the limitations of grid-size or map-base dependencies typical of more common interpretations of Lagrangian data.

ACKNOWLEDGEMENTS

The author would like to acknowledge many helpful discussions and editorial suggestions by Renn Hanson and D. L Payton that have contributed to this work

Table of Contents

BACKGROUND

Lagrangian trajectory models provide a flexible framework for studying a wide variety of pollutant distribution problems. They typically model the pollutant as an ensemble of LE's (Lagrangian elements or particles) and handle the mixed scale problem (releases are initially very localized, yet eventually become widely dispersed) without significant numerical dispersion. A limitation of this approach is that the model output is a set of LE locations, which when viewed graphically appears like a swarm of bees. This gives a qualitative indication of the forecast location of the pollutant, but what is usually desired is a quantitative Eulerian density field. The Lagrangian-to-Eulerian transformation is not straightforward in realistic geophysical domains, but satisfactory results can be be obtained for floating distributions using tessellation methods based on Thiessen polygons (Galt,2011). For floating distributions of LE's the Eulerian densities are dimensionally given as $mass/km^2$, but this approach breaks down when applied to stranded or beached pollutants distributed along a shoreline where the Eulerian density will be dimensionally represented as $mass/km$. It is obvious that a different sort of analysis is required for the “floating” vs. “beached” ensembles of LE's. The remainder of this note addresses the Eulerian derivation for the “beached” ensemble of LE's.

Assume that we are given a collection of particles that are randomly distributed along a linear feature (a shoreline) as a Lagrangian data set which have associated position information $r(x_i, y_i)$ indexed over

i= 1,2,3... . It is often useful to consider the equivalent Eulerian density of this distribution which will be dimensionally given as mass/length. The algorithmic transformation that this requires can become quite complicated when dealing with complex and possibly fractal shorelines. It is therefore desirable to investigate the possibility of robust alternate analysis methods that will not depend on assumptions about an underlying map base. One such option would be to consider cluster analysis, such as is commonly used in neural network and AI applications (Xu and Wunch II,2009).

We will start by considering a uniform but random collection of point masses. We then seek a linear metric that represents the neighboring mass in proximity to any individual point mass. In some sense a measure of the "clustering" of mass around that point. One such trial function is the Gaussian kernel where the local length-metric (dist[i]) is given by:

$$dist[i]=\int_0^\infty e^{-k^2\|r-r_i\|^2}dr=l_k \qquad (1)$$

And with the (x,y) coordinates of mass points used to represent the norm:

$$\|r-r_i\|^2=\sum_i\left((x-x_i)^2+(y-y_i)^2\right) \qquad (2)$$

As r → ∞ in the definite integral seen in equation(1) approaches a limit. We can represent the integral numerically using a point collocation technique so that equation(1) can be written:

$$dist[i]=e^{-k^2\sum_j((x_i-x_j)^2+(y_i-y_j)^2)}=\frac{\sqrt{\pi}}{2k} \quad (3)$$

This is a linear measure that selectively weights nearby point mass objects. The value k is seen to scale the assumed uniform density of mass in the vicinity of the point mass. As an example, if the position data is given in kilometers with k=1 the "nearness kernel" would imply a "standard uniform distribution" of (unity/per kilometer). We can define that as our reference field l_0

$$l_0=\frac{\sqrt{\pi}}{2} \quad (4)$$

Now consider what would happen if we were to numerically evaluate the dist[i] value for each mass point using equation(3).

$$dist[i]=\frac{\sqrt{\pi}}{2k_i} \quad (5)$$

This would provide a value of k_i which is inversely proportional to dist[i] and normalizing with equation(4) we will have:

$$dist[i]=\frac{l_i}{l_0}=k_i^{-1}(kilometers) \quad (6)$$

From this we see that applying equation(3) to each point mass in the Lagrangian set provides a length-metric (in terms of the value k)

that scales the "nearness" of neighboring masses by scaling the Gaussian kernel. As k increases the length-metric decreases like the reciprocal of the local assumed uniform mass distribution. This then is the cluster length scale we were seeking and a proper representation of the Eulerian density field at the location of any Lagrangian point will be the particle mass divided by the scale distance

$$\sigma_i = density_i = \frac{m_i l_0}{dist[i]} = m_i k_i (mass/km) \tag{7}$$

From continuity considerations, we require that the integral of density over the domain must add up to the total mass:

$$\int \sigma_i \delta(m_i) = M_{beached} \tag{8}$$

The will not be a normal Riemann integral, but the measure can be represented as a type of Stieltjes integral (Niamark,1968) where the differential length associated with each mass particle will be: $\delta(m_i) = l_i(local\,length\,unit)/l_0(scaling\,factor\,for\,length)$ Using this and equation(6) gives:

$$\int \sigma_i \delta(m_i) = \sum \sigma_i \frac{l_i}{l_0} = \sum m_i k_i \frac{l_i}{l_0} = \sum m_i = M_{beached} \tag{9}$$

Which confirms the result that the sums of the particle masses conserves total mass. In addition, the sums of the product of local densities times the linear-metric normalized by total mass is less then or equal to unity. This means $l_i \sigma_i / l_0 M$ could serve as a

surrogate for probability density when considering the position data of each particle as an independent event. In particular the probability density of the state represented by particle i will be:

$$probability\ density = p_i = \frac{l_i \sigma_i}{(\sum l_0)(M_{floating} + M_{beached})} \qquad (10)$$

And we note:

$$0 \leqslant p_i \quad \text{For each i}$$
$$\text{and} \qquad (11)$$
$$\sum p_i \leq 1$$

Where the < condition in the second line of equation(11) is due to the fact that analyzing the beached particles generally does not include the full potential of beaching. Some of the particles may still be floating and could eventually strand on new (or previously oiled) segments of the beach. As time goes on the beached fraction of the potential masses used in equation(9) will become a larger fraction of the total particle masses used in equation(10) and eventually, if all the particles are located on the boundary, the sum of the probabilities will be unity.

If the Lagrangian fields represented probabilities rather than mass particles then the resulting Eulerian calculations would still represent probability densities where total mass would be replaced by unity (certainty) in equation(10). Section three of this note will explore the probability implication of equation(10) in more detail.

ANALYSIS EXAMPLE

As an example of this approach I will demonstrate using the output from a 24 hour GNOME trajectory run in Burrard Inlet (Tsleil-Waututh Nation,2015 appendix 2) with a hypothetical release of 8000 particles on a flood tide in the inner harbor between the 2nd Narrows and Burnaby Narrows. The initial movement is due to a flood tide into Burnaby Narrows, but after two tidal cycles the LE's are distributed pretty much along the length of Burrard Inlet. Ebb and flood tides, variable winds and a 18 hour re-float half life all contribute to the distribution and, of the original 8000 particles over 7000 are beached. A plot just showing the distribution of beached particles is as follows:

Figure 1 – Plot of all beached particles

The detailed analysis goes through the following steps:

1) for each LE the cluster distance eq(5) is calculated $l_i = k_i^{-1}$

2) the mass density is scaled giving σ_i

The beached particles are replotted with the radius of the LE's proportional to the square root of the density subject to a minimum of 1 pixel to ensure the full distribution is included. The plotted size of the splots is held constant during a zoom of the figure so that a detailed view of the Burnaby Narrows still presents the heavily oiled segments, but does not obscure the details of the underlying map.

Figure 2 – Plot of all beached particles scaled to Eulerian density

This gives a very different visual representation of the distribution of beached particles. The clustering of particles near the initial release sites is clearly seen whereas in the standard plot (showing all of the LEs plotted independent of clustering) this facet of the distribution is easily missed. In addition, a quantitative measure of the local oil density is available for each hit location.

It should also be noted that the beached-oil distribution can not be inferred from floating-oil distribution and the relationship between these two distributions is complex depending on the wind history and oil retention characteristics of the shoreline. For comparison the distribution of the approximately 1000 remaining floating particles is shown below:

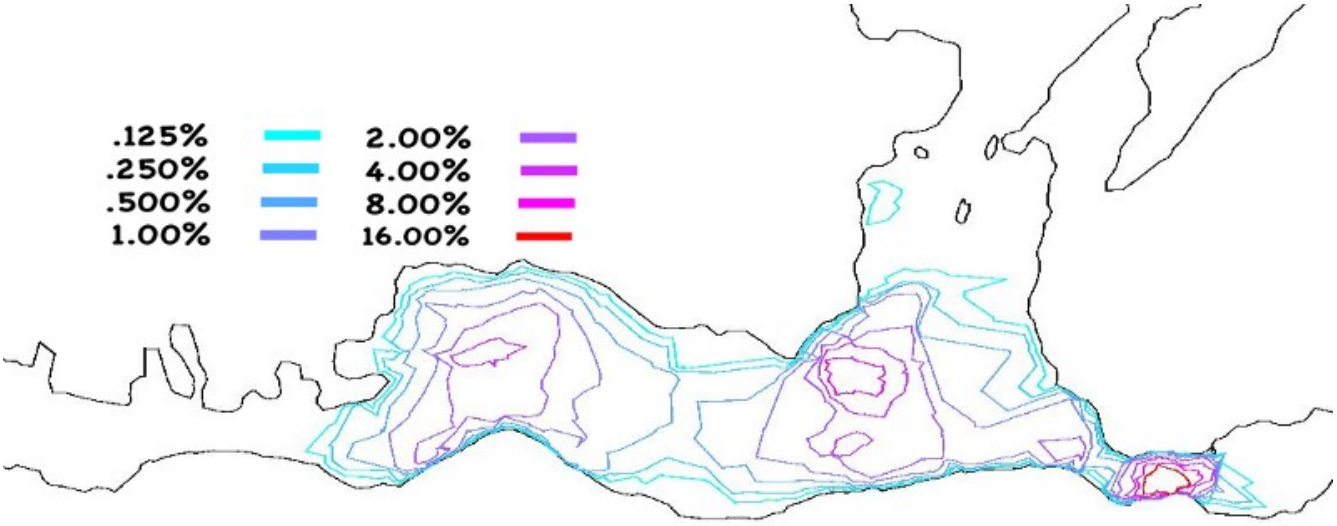

Figure 3 – Distribution of floating particle density ($mass/km^2$)

An even better understanding of the beached particle distribution can be obtained by looking at a length-metric analysis graph of the particles. This graph is constructed by calculating the length scale for each particle and then sorting them in descending order. Plotting these values with particles along the horizontal axis and scale distance up the vertical axis results in the following plot with the blue region representing the sorted scale metrics:

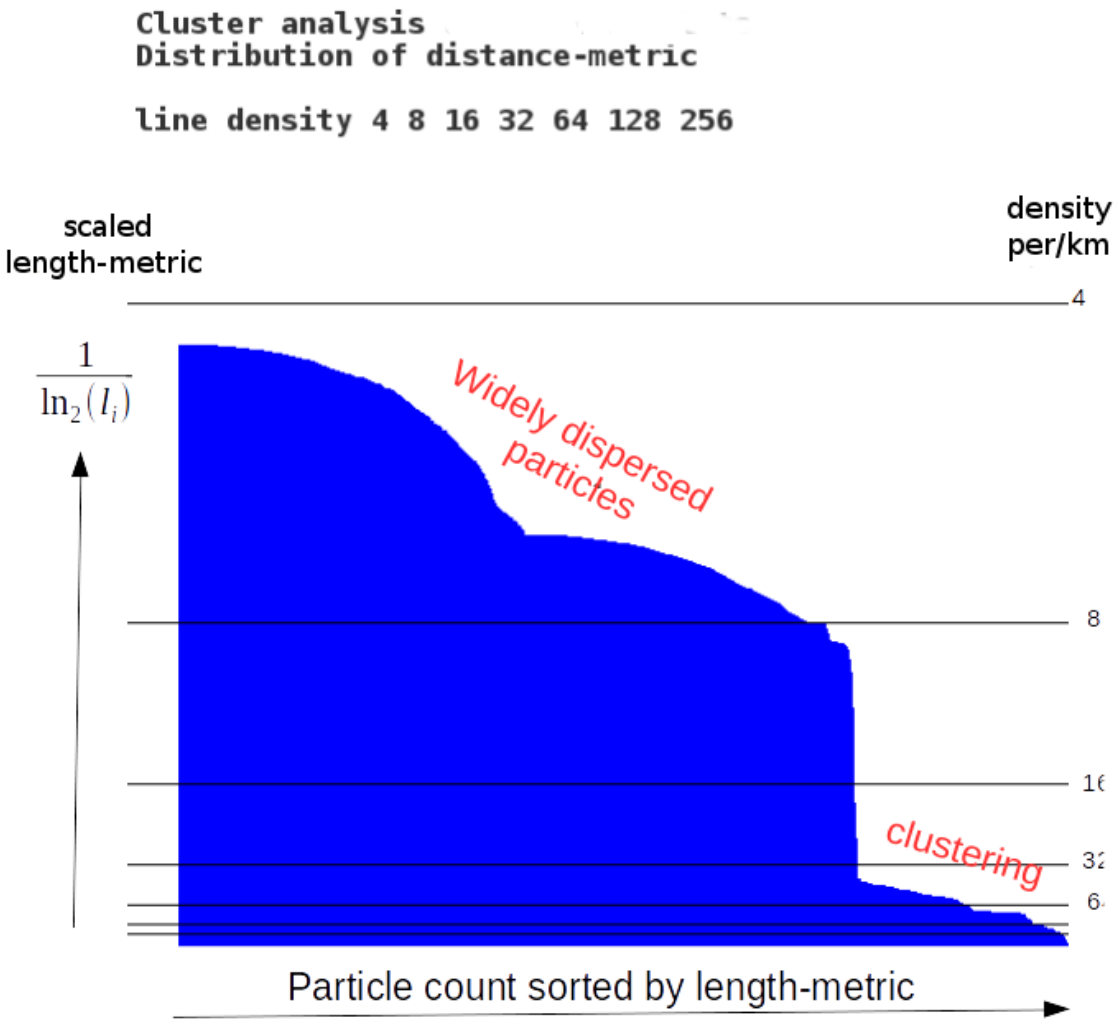

Figure 4 — Plot of the sorted length-metric where the vertical axis represents $1/\ln_2(l_i)$ and the horizontal axis is particle index.

Also drawn on this figure are horizontal lines that indicate the absolute scale values of linear density where the upper line gives the standard unit/km that is below the minimum value of the data set and each subsequent lower line represents an increase of a factor of 2 in the density. This shows that three quarters of the particles

(the left hand side of the plot) are generally scattered in a slowly increasing density between 4 and 8 particles/km with no particular evidence of clustering. At that point a sharp break in the curve shows clustering taking place with approximately a quarter of the particles in clusters where densities are on the order of > 32 particles/km. Cluster densities then rise steadily to values exceeding 200 particles/km.

CLUSTERING AS INFORMATION

When considering that each particle is associated with it a length-metric it is possible to consider the “information content” of the trajectory model that provided the output. This approach is well known in communication studies and was initially developed by Shannon's information theory work at the Bell Labs (Shannon and Weaver,1963). The basic ideas center around signal “uncertainty” and “entropy” and, since its introduction, has seen applications in many fields such as Artificial Intelligence, data compression and cryptanalysis.

To introduce these ideas, consider a diffusive model process that completely spreads out a cluster of particles so that associated with each element is a constant length-metric l_0 so that the entire group of particles covers a distance of Nl_0 . In this case, the uncertainty in location of any individual particle will be (at least statistically) represented by l_0 In some way l_0 is related to the entropy of the particle.

As an alternative, consider a case with the same diffusive process but restricted in time or by physical processes (kinematic – like shorelines, or dynamic – like currents and winds) so that clustering remains present and particles don't spread to a uniform final state. In this case the length-metric associated with the clustered particles will be l_i which will on the average be smaller then l_0 . This means that the particle in question is more closely confined.

From a communication of information point of view the statistical difference in bits required to represent l_0 vs. l_i provides a data compression and can be thought of as information supplied by the model. For a collection of particles the sum of the bits (and fractions of bits) gained by all of the clustering information is the model's intelligence. As particles spread out and their positions become less certain the "entropy" of the solution increases. Information entropy like chemical entropy is always referenced to some base state (0 degrees Kelvin — for chemistry) and uniformly random distributions for information. To summarize: If we had no information, particles could be anywhere with a uniform probability implying a base state entropy. A model provides clustering information and implies a lower state entropy. The difference between these two is a quantitative measure of information provided by the model. The major contribution provided by Shannon's pioneering work was his quantifying the form that information entropy functions must take.

In a message made up of digital signals for which each signal has a known probability p_i , Shannon defined the total "entropy" of the message as:

$$H = -\sum_i p_i \ln_2 p_i \qquad (12)$$

As in the case of thermodynamics the entropy is always referenced to some base factor (such as absolute zero in chemical studies) and the interest is always focused on changes in the value of H rather then its absolute value. The changes in the value of H determine how a process is progressing toward uncertainty. From a qualitative point

of view this is easy to understand in dispersive models. A tight grouping of particles (low entropy) spreads out leading to a weaker understanding of where individual particles are (higher entropy). If the domain is bounded the maximum entropy will be a state where all the particles are at a uniform density filling the domain. Shannon proved as a formal requirement of equation(12) that this would be the final (equivalent to absolute zero) maximum entropy state. With this much as background we may consider applying equation(12) to what we discovered from equation(10) of this note.

Representing the sum of all the probability densities as represented by equation(10) we get:

$$\sum p_i = \frac{\sum l_i \sigma_i}{(\sum l_0)(M_{floating} + M_{beached})} \qquad (13)$$

This clearly shows that the individual particle probabilities in the numerator are normalized by a cumulative span in the denominator. It is a simple function of the assumed length scale used for the reference. If we were in a bounded domain it would be the domain dimensions divided by the number of particles.

In our study case we used l_0 = 1 km as a reference scale. As a measure of our final uncertainty this would probably be an under estimate if we were modeling the North Pacific an over estimate if we were modeling a small lake.

As an alternative to an arbitrarily chosen length reference we could let the model domain determine a self describing length scale defined as the mean of all the $l_i's$:

$$l_m = \frac{\sum l_i}{N} \qquad (14)$$

This value would be self scaling in some respect, but implies that the final ground entropy state was nothing but a redistribution of the particles over the sections of the shoreline where it happened to be at the time of analysis. The probabilities would sum to unity and the "hit space" would be assumed fixed.

A better approach might be developed by going back to the trajectory model and determining how much actual shoreline is available as a target space for the stranding of particles. This would set the scaling length to:

$$l_{mod} = \frac{\sum ModelShorelineSegments}{N_{floating} + N_{beached}} \qquad (15)$$

The final base entropy would then be defined as one particle per l_{mod} length of shoreline (assuming all the particles beached). This may seem like we are going back to an underlying map (potentially fractal) problem, but this is not really the case. The model has a unique shoreline (defined by the map.bna file in the case of GNOME) and its segments can be determined with a simple algorithm. The number of available particles is also available from the model output, so the functional length scale defined by equation(15) is available at the time of analysis.

Using l_m or l_{mod} in place of l_0 in equation(13) would give a different value for relative entropy, but once again the real issue is the observed changes in entropy not the absolute value relative to some base. These are alternate ways of looking at the same problem. It is certainly likely that the sum of p_i values associated with the

mass points well be less then unity (some of the LE's are still floating) and there are assumed to be available states which are presently empty (segments of the shoreline that have not been oiled). This is equivalent to the statement that the particles are not uniformly distributed yet. Applying these equations to our example case results in the data shown in the following table:

---

min nearest neighbor 1.26E-4 max nearest neighbor 0.183
maximum length-metric in analysis 0.2215 maximum line value 4
total number of beached particles 7404
total hit space 332.71 km (from model bna map)
lmod (unit hit space) 0.0415
unit Base Entropy 0.001619 Total Base state entropy 12.95
total hit probability 0.1152
present entropy is 1.805
entropy ratio 0.1393

Table 1 – Quantitative Analysis of Clustering Data, Probability Density, and Model Entropy for test case.

Before leaving the discussion of information content associated with probability densities and entropy we should note that in the previously mentioned reference (Xu and Wunch II,2009) a nonlinear Principle Component Analysis is described. This procedure allows the kernel results in dist[i] to be transformed as a non-linear mapping into a covariance matrix which then makes it possible to obtain EOF eigenvalues and eigenvectors using standard EOF methods. The eigenvectors provide a graphic representation of dominant clusters and the eigenvalues quantify how much model variance they contain.

TRAJECTORY MODEL EVOLUTION, CLUSTERING AND ENTROPY

The initial aim of this note was to consider a linear metric to represent the Eulerian density of a Lagrangian distribution known to be confined to a general curvilinear domain. Introduction of a Gaussian kernel approach for clustering led to the ability of deriving a shoreline density (mass/length) which did not include considerations of a background map or introduction of a raster grid. Several graphical examples demonstrated that these computational methods applied to the output of a Lagrangian-oil-trajectory model's beached (or stranded) data provides useful presentations in an Eulerian form. Lagrangian to Eulerian analysis of model output representing floating oil using tessellation with Thiessen polygons provides mass/length-squared representation (Galt and Hanson,2015). Although the computational methods are quite different, this pair of analysis techniques provide a powerful methodology to process and present Lagrangian oil trajectory model output.

The availability of Eulerian density data: ($mass/km\ for\ stranded$) and ($mass/km^2\ for\ floating$) pollutants enables us to define probability density functions for the distributions predicted by the oil trajectory model. From these it is possible to consider the model output as a data channel or stream of information with a well defined "entropy". The time development of the model entropy is an integral measure of what is going on with the overall model dynamics. It is a characterization of the regional dynamics. It describes how much the

model can tell us and how rapidly it's predictive power fades. Consider the following:

The ocean surface is a turbulent and chaotic environment. Particles floating on it disperse. This can be thought of as a destruction of clustering. Initial clusters spread. The entropy representing the probability densities of particle locations will increase monotonically. If no other physical process were active this increase in entropy would continue to rise and eventually asymptote when the particles are uniformly distributed over the domain. This is not a particularly interesting, but well defined entropy curve.

Now we can reconsider the problem and introduce physical processes. These will be in the form of winds and currents (“movers”) and shorelines intercepting the resulting “moving particles”. These processes can represent anything. Those that are simple translations won't change the time dependence of the entropy curve, but virtually any other behavior will. Shears and divergences will enhance dispersion (lessen clustering) while convergences and stranding on shorelines will represent anti-dispersion (clustering). All the effects of these processes will be reflected in the time dependent curve of the model entropy. Regions where shear and divergences dominate will show a relatively steeper rise in entropy which will indicate pollutants will be more difficult to locate and encounter. On the other hand convergence and beaching, leading to clustering, will cause a decrease in the rise of entropy which may even plateau or decrease. This will indicate trapping of pollutants and time horizons when higher concentrations will present themselves.

We can expect that the general time dependent shape of entropy curves

will characterize the fundamental dynamics of the region. The strong "convergences" in Cook Inlet and at the fresh-water interface along coastal Gulf of Mexico should lead to clustering behavior. Sea breeze regions may also show periodic clustering enhancements due to beaching. Divergent flows over shoals should show divergent behavior and destroy clustering. Time dependent entropy analysis of regional trajectories should provide additional understanding about how much information we might expect from model studies, how long they are likely to provide useful information without requiring new data assimilation, and particular time windows when trapping or cluster formation might occur.

REFERENCES

Galt,2011. Triangular Tessellation Documentation – Genwest Technical Publication 11-001, Genwest.com/genwest_publications

Galt and Hanson,2015. Description of a fast general routine to process Lagrangian Point into Eulerian density distributions. Presentation at LAPCOD IV (Lagrangian Analysis and Prediction of Coastal Dynamics) meeting, Winter Harbor, MA abstract published.

Shannon and Weaver,1963. The Mathematical Theory of Communication Univ of Illinois Press, Urbana and Chicago 125p

Tsleil-Waututh Nation,2015. Assessment of the Trans Mountain Pipeline and Tanker Expansion Proposal, Appendix 2. Treaty,Lands & Resources Department, Tsleit-Waututh Nation.

Xu and WunschII,2009. Clustering IEEE press, Wiley and Sons INC 358p