Automated Quality Control of Tropical Cyclone Winds Through Data Mining

H. Nicholas Carrasco
Cooperative Institute for
Marine and Atmospheric Studies
University of Miami/RSMAS
NOAA/AOML/Hurricane Research Division
Miami, FL 33149
Email: Nick.Carrasco@noaa.gov

Mei-Ling Shyu
Department of Electrical and
Computer Engineering
University of Miami
Coral Gables, FL 33124
Email: shyu@miami.edu

Abstract

The analysis of tropical cyclones (TC) depends heavily on the quality of the incoming data set. With the advances in technology, the sizes of these data sets also increase. There is a great demand for an efficient and effective unsupervised quality control tool. Towards such a demand, data mining algorithms like spatial clustering and specialized distance measures can be applied to perform this task. This paper reports our findings on the studies on utilizing a density-based clustering algorithm with three different distance measures on a series of TC data sets.

1 Introduction

Tropical Cyclones (TCs) occur throughout the world's oceans and are monitored globally by aircraft-, space-, and earth-based observing systems. Advances in computing and communications have made it possible to obtain these observations from around the world. To properly understand the life-cycle of a TC, hurricane specialists and meteorologists rely on observations, models and analyses of the environment. These models and analyses combine observations from the above mentioned observing systems collected in near real-time. However, the model results and analysis can only be as accurate as the data that is collected and used. Hence, quality control (QC) of the observation points is critical to the preparation of data for analysis.

QC is a time consuming task performed subjectively by hurricane specialists and meteorologists, and is largely based on personal experience and trust in the various observing platforms. As technology advances, so does the quantity of the observations that must be quality controlled. The quality of remote sensing platforms such as Doppler RADARs and satellite are quickly improving. Improved wind retrieval algorithms allow for observations to be measured at higher resolutions thus increasing the number of observations that can be obtained. In order to keep up with the growing demands of larger data sets, data mining routines, such as clustering, look promising in automating the quality control of TC meteorological observations. Spatial clustering schemes can be used to quickly and effectively detect noise in the data.

This paper is organized as follows. Section 2 briefly gives an explanation of the data. In Section 3, how the clustering technique can be used to facilitate the QC of the observations is discussed. The DBSCAN clustering algorithm and the issues involving in the implementation with TC data are also presented in Section 3. Section 4 gives the results. Finally, conclusion and future directions are discussed in Section 5.

2 Current State of TC Data

The global meteorological observing system has been evolving for 300 years, but the current real-time system is a product of the technology and communications in the 19^{th} century [2]. Today's global network of meteorological observing systems includes observation stations on the Earth's surface both on land and sea. Remote sensing platforms including RADARs and satellites in space. There are also various airborne observing systems around the world, such as weather balloons and reports from commercial aircrafts.

The National Oceanic and Atmospheric Administration (NOAA) operates research and reconnaissance

missions in conjunction with the 53rd Weather Reconnaissance Squadron of the 403rd Wing, Air Force Reserves, into North Atlantic and East Pacific tropical cyclones [8, 14]. These aircrafts carry an array of in-situ and remote sensing instruments which include wind anemometers as well as thermometers and barometers, precipitation probes and several Doppler RADARs located at key positions of the aircraft. One cutting-edge remote sensing device used is the Stepped-Frequency Microwave Radiometer [23] which can estimate winds near the surface. GPS Dropwindsondes are ejected from the aircraft into the storm and they can provide in-situ measurements from the aircraft flight-level all the way to the ocean surface generating useful vertical profiles [6]. All these measurements are considered by the meteorologists when studying TCs for research and forecasting.

NOAA's National Weather Service (NWS) and its National Hurricane Center (NHC) in Miami, FL have the responsibility to monitor and inform the public of a TC's status in the North Atlantic and East Pacific Oceans. The NHC hurricane specialists use all of the above mentioned observing platforms in conjunction with Numerical Weather Prediction (NWP) models and objective analysis products to make decisions regarding the future track and intensity of a TC, and to declare evacuations in the event of a landfall. The NWP and analyses used depend heavily on the quality and quantity of the observations assimilated. The surface wind analysis used by NOAA's Hurricane Research Division (HRD) of the Atlantic Oceanographic and Meteorological Laboratory (AOML) in Miami, FL, is an objective spline analysis [15,16]. The HRD Spline Analysis (HSA) attempts to generate a uniform estimate map of the current state of a TC based on a quality controlled set of observation points.

While technology has improved some of the quality, it has also increased the quantity of observations. Manual QC of the data has become an overwhelming task. *Gross error* checking can remove some points, but the bigger challenge occurs when there are conflicting neighboring data points. After many years of studies, the various observing platforms have been assigned relative weights [7, 17–20], which are used by HSA in an attempt to handle these areas with conflicting data points.

3 Clustering TC Data

As mentioned in Section 2, a combination data from space-, air- and surface-based observing platforms are assimilated into NWP models and analyses. The two main issues related to their quality are the accuracy of the measured values and the spatial coverage of the data points. While there is little to be done about the areas of low density (other than to add more stations and newer observing systems), the efficient and effective QC of the data in the areas of higher density is vital to the estimation of the current state of the environment (see Figure 1 for a sample data coverage plot). Currently, other than gross error checking, some models and analysis systems have incorporated other QC techniques such as *buddy checks*, commonly known as *nearest-neighbor* checks. However, as well known by the data mining community, the efficiency of these checks are heavily affected by the size of the data set.

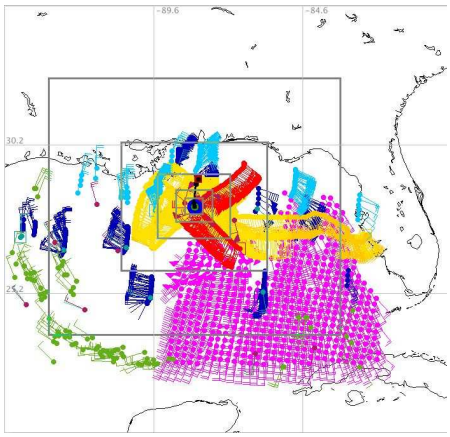


Figure 1. Sample of data coverage of non quality controlled TC data from Hurricane Ivan (16 September 2004 0130UTC)

H*Wind, a tool developed by HRD, allows the meteorologists to visually interact with data from the various available sources, and manually QC the observations. However, as mentioned earlier, this is a time consuming and subjective task. The need for an accurate estimation of the current environmental state requires a timely and efficient QC method. In addition, this process should not depend heavily on human interaction, due to its subjective nature.

Recent developments in the field of data mining and clustering in particular, provide a good starting point for the development of an unsupervised QC method for the meteorological observation data sets. Since these observations are collected on a global scale, their spatial nature makes spatial clustering routines stand out as a viable tool.

Clustering is the unsupervised classification of patterns found within data [9]. There are basically two types of clustering algorithms, partitioning and hierarchical [10]. *Partitioning algorithms* attempt to divide a

database D into k clusters based on its relationship to the cluster's center (*k-means*) or to some central member of a cluster (*k-medoid*). *Hierarchical algorithms* attempt to generate clusters by taking either the whole database and decomposing it one level at a time, or by taking each point as an individual cluster and merging them to form larger and more descriptive clusters.

In the area of *spatial clustering*, many algorithms have been developed over the years [3, 4, 11, 13, 22, 24, 25]. Spatial clustering algorithms attempt to find the clusters over a geographic space. Several spatial clustering algorithms also have been developed with the intent to handle the existence of noise in data sets. The primary issue with meteorological data sets is that not all points that would be classified as noise are really noise.

Most clustering algorithms tend to label data points that do not fit into a cluster, as *unclassified* points or noise. As just mentioned, in the meteorological observation network, a majority of the world is not constantly monitored. Satellites provide a good source of observations in the areas of low coverage, but polar-orbiting satellites only measure points approximately once every 12 hours. Consequently, there tend to be observation points without *neighbors*. The drifting buoy and volunteer ship reports collected can be located at any random location throughout the world's oceans. While these reports may not always accurate[1], they are often the only source of data in key areas. Also geostationary satellites are limited to retrievals in the visible and infrared spectrum due to their relatively higher orbits causes retrieved observations to be spread out [12]. To consider all data points without neighbors as noise, would leave many areas completely void of any measurement. This can be detrimental to analyses and NWP models. When implementing a clustering algorithm, this must be taken into account. At the same time, clustering makes it relatively easy to find noise in the areas of higher density.

3.1 Selected Clustering Algorithm: DBSCAN

When choosing a clustering algorithm, there are many details that depend on the data set involved. Most clustering algorithms, being unsupervised, use assumptions taken from the data set to help define partitions, and consequently the resulting clusters may vary [5]. The same algorithm using different assumptions about the data and different input parameters can result in completely different clusters. Based the literature [9, 11], DBSCAN was selected as the algorithm for the task at hand. DBSCAN (Density Based Spatial Clustering of Applications with Noise) [3] is a density-based clustering algorithm. Other algorithms researched included grid-based algorithms like STING [24] and WaveCluster [22], and other partitioning algorithms like CLARANS [13].

The initial algorithm selected was CLARANS for its speed through the use of a randomized search, but as discovered in further readings, while designed for large data sets, it does not perform as well as many other algorithms. The grid-based algorithms were evaluated based on their speed, but not selected due to the fact that they do not effectively account for points on the border between two clusters. Based on the average data coverage and the known physics of a TC, density-based algorithms like DBSCAN and others like DBCLASD [25] were evaluated.

The DBSCAN algorithm can be found in [3, 4]. In summary, DBSCAN first determines if a point is a *core point* or a *border point*, and then finds all *density-reachable* points from a core point and classifies it as a cluster. If a point is not reachable, it is labeled as *noise*. It takes two values as input parameters, *Eps* and *MinPts*. MinPts is the minimum number of points it takes to be a cluster, and also determines whether a point is a core point or a border point. Based on Ester et al. [3, 4], the value of MinPts can be fixed at four points and choosing anything larger did not produce significantly different results. The value of Eps is the "distance" from a point that another point can still be considered as a neighbor. They also describe an objective way of determining a value for Eps, but for the purposes of this paper, this value was determined based on the discussions with hurricane meteorologists and specialists.

3.2 Implementation Issues

As already mentioned, clustering results depend heavily on the assumptions taken from the data set. These assumptions help when designing and improving a distance function. Distance functions, while they can be very generic, when tuned for a data set, the more accurate the results will be. Attributes commonly available to meteorological observations include: winds, temperature, humidity, and pressure. Each measurement comes from an independent measuring devices with the exception of satellites which derive their values based on other observable quantities. While there are some relations between measurements that can be derived, the measurements do not necessarily depend

[1]Volunteer ship reports are taken by government and commercial sailors. These sailors are not always scientists or meteorologists, and while they do follow a set of standards, the quality of the observations can be inconsistent.

on one another. In this paper, the attributes, unless otherwise specified, will be focused on location, time, wind speed, and wind direction.

As seen in Figure 1, the density of the observations varies throughout the domain. Like most clustering algorithms, DBSCAN labels all points without a significant number of neighbors to be noise. However, it does provide an excellent support for detecting noise within the areas of high density. The issue of neighbor-less points also depends on how one implements the distance function.

Spatial clustering schemes like all clustering schemes, including DBSCAN, are heavily dependent on the *distance* functions. In general, spatial clustering schemes use Cartesian coordinates like Mercator latitudes and longitudes to represent location and simple distance measures like Euclidean and Manhattan. As already mentioned, the use of Cartesian coordinates and simple Euclidean distance measures can lead to a majority of the observation points to be labeled as noise or with no neighbors.

When there is a significant amount of information known about the data set to be clustered, even these simple distance functions can be tuned. In the case of TCs, there is much known about the physical structure. As discussed in [1, 15, 16, 21], a key feature of TCs is their cylindrical nature. Winds within a TC revolve around the *center of circulations*[2]. Based on this knowledge, an implementation can take into account these structural features. Another key feature about TCs is that it is relatively symmetric in nature, if the wind direction of an observation on one radial is of a particular value, the wind direction values on the opposite radial should approximately be the inverse. It has also been noted that TCs behave slightly different in each quadrant, and that the observations within one quadrant should be significantly more similar to each other than to the observation in another quadrant.

With this knowledge, distance functions can be tailored and tuned. Taking into account the cylindrical nature of a TC, using the Cartesian coordinate system can be replaced with a polar/cylindrical coordinate system using the storm center as its reference or origin. This polar coordinate system changes how the observation density looks. When determining the spatial difference between two observations, one can compare the radial distances from the center or in conjunction with its angular differences. The radial distance from the center allow the observations with no Cartesian neighbors to be compared with other observations with a similar distance to the center but in any quadrant of the TC. Adding the angular difference quadrant features are more prevalent.

[2]The *center of circulation* is the *center*, *vortex*, or *eye* of the storm. The storm is said to revolve around this point

4 Result Analysis

As mentioned earlier, DBSCAN was implemented with the option of three distance measures based on different coordinates systems. The three methods were: *Euclidean*, *radial*, and *polar*. For the radial and polar methods, a center was selected based on the *storm track* positions from NHC and corresponds to the date and time of the analysis. As mentioned in Section 3.2, the wind direction is rotated as the points are looked at around the storm. Therefore, the wind direction was not considered when comparing the radial distances from the center. Due to the idealized symmetrical nature using the radial distance method, spikes in wind speeds could be detected and removed as noise. Similarly, when performing the polar distance method, using the angular distance makes it possible to remove the observations with erroneous wind directions. For all of the mentioned methods, the maximum difference values (Eps values) used by DBSCAN were taken based on the discussions with the meteorologist and specialist.

In order to check the validity of the results, each distance measure was applied to a series of TC data sets. The selected data sets are all from the 2004 Atlantic hurricane season. They include Hurricane Frances on September 4^{th} at 2230UTC, Hurricane Ivan on September 16^{th} at 0130UTC, and two cases from Hurricane Jeanne, one on September 17^{th} at 0730UTC and the other on September 26^{th} at 0130UTC. The times selected correspond to the times of available operational analyses for comparison. These operational analyses were done in real-time and were quality controlled by hurricane meteorologists. These operational analyses values were considered to be the "truth."

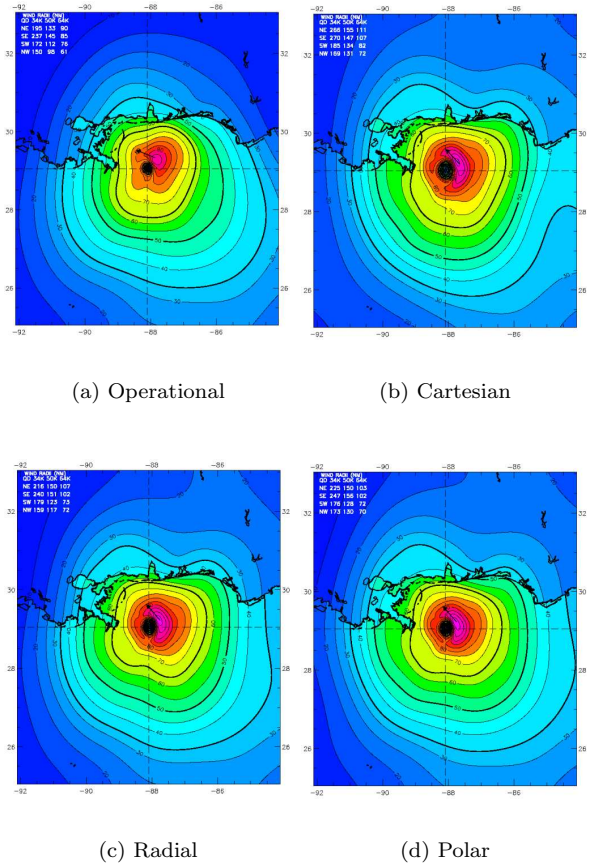
The mean radii %error per quadrant are shown in Table 1. These values were derived by taking the %error for each case and then averaging the values together for each quadrant. The %error is used as a method to normalize the results between individual cases. While all hurricanes have a similar structure, they usually have varying dimensions. The wind radii, as defined by the NHC, are the maximum radius from the center, where the winds of the denoted wind speeds can be expected. The standard wind radii used by NHC are 64kt (hurricane force), 50kt, and 34kt (tropical storm force).

As shown in the table, the errors match what are expected. What was not expected was that the results using the Euclidean distance with Cartesian coordinates

Table 1. Mean %error for wind radii in each quadrant (%error = |x-y|/y)

	64kt Wind Radii				50kt Wind Radii				34kt Wind Radii			
Measure	NE	SE	SW	NW	NE	SE	SW	NW	NE	SE	SW	NW
Operational	0.55	0.28	0.16	0.67	0.31	0.25	0.36	0.60	0.23	0.15	0.32	0.20
Cartesian	0.67	0.32	0.31	0.24	0.36	0.20	0.42	0.45	0.29	0.20	0.40	0.21
Radial	0.17	0.23	0.15	0.14	0.10	0.21	0.18	0.16	0.33	0.60	0.12	0.06
Polar	0.15	0.22	0.13	0.15	0.36	0.40	0.20	0.21	0.52	0.66	0.11	0.54

had a higher %error on average than those not using any clustering method for QC. Both the radial and polar did fairly well, with the note that the polar method performed fairly better closer to the center. It must be noted that the further away from the center the winds were, the less QC was necessary. However, more work must be done to better handle the winds closer to the center as can be seen in Figure 2 of the wind contours from HSA. The darker region in the center, denoting more wind contours, in all three methods seems to be significantly larger than that of the operational analysis. This could be a result of the rapidly changing wind directions near the center of circulation, which would explain the better handling near the center of the radial measure. As mentioned earlier, the radial measure does not evaluate the wind direction.



(a) Operational (b) Cartesian

(c) Radial (d) Polar

Figure 2. HSA Wind contour plots

5 Conclusion and Discussion

In this paper, the basic use of different distance measures and coordinate systems for quality controling TC observations has been discussed. As expected, the resulting clusters and therefore the resulting noise excluded due to the QC, varied based on the used method. While closer to the center, the radial and the polar methods outperformed the Euclidean-Cartesian method. In addition, it was noticed that as the observations moved further away from the center, almost no QC was needed. From the discussions with the meteorologists, this seems relatively true based on the time spent manually QCing data on the perimeter of the analysis domain.

Based on the results, there are several other techniques that may be useful. One technique that may help is to use a more precise distance measure based on the polar coordinate system, the *arc distance* should be evaluated against the polar method used here of taking the radial and angular differences. Another possible technique for handling the structure of the TC may be to vary the way one looks at the wind values. Wind values are typically viewed in terms or magnitude and direction, but can also be viewed as vector components. This should not drastically change the results. However, wind values can also be looked at in terms of the radial and tangential components. This method may allow the simple Euclidean-Cartesian method to take into account the cylindrical shape. Another topic for further study is to possibly dynamically determine in which cases to use different distance measures. What is not accurately shown in the presented results was that for TC events consisting of primarily low winds (<50kt), the Euclidean-Cartesian method outperformed both the radial and polar methods almost two-fold. Furthermore, it may be necessary to use a combination of distance measures to properly QC the TC data.

6 Acknowledgments

This research was carried out in part under the auspices of the Cooperative Institute for Marine and Atmospheric Studies (CIMAS), a Joint Institute of the University of Miami and the National Oceanic and Atmospheric Administration, cooperative agreement #NA17RJ1226. "The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency."

Special thanks to Dr. Mark Powell, Shirley Murillo, Sonia Otero, Jason Dunion and the rest of HRD and the H*Wind team.

References

[1] K. H. Bergman and T. N. Carlson. Objective analysis of aircraft data in tropical cyclones. *Monthly Weather Review*, 103(5):431–444, 1975.

[2] R. Daley. *Atmospheric Data Analysis*. Cambridge University Press, Cambridge, MA, 1991.

[3] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, Portland, Oregon, August 1996.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. Density-connected sets and their application for trend detection in spatial databases. In *Proceedings of 3nd International Conference on Knowledge Discovery and Data Mining*, pages 10–15, Newport Beach, CA, August 1997.

[5] M. Halkidi and M. Vazirgiannis. A data set oriented approach for clustering algorithm selection. In *Proceeding of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 165–179, Freiburg, Germany, September 2001.

[6] T. F. Hock and J. L. Franklin. The NCAR GPS Dropwindsonde. *Bulletin of the American Meteorological Society*, 80(3):407–420, 1999.

[7] S. H. Houston and M. D. Powell. Surface wind fields for florida bay hurricanes. *Journal of Coastal Research*, 19(3):503–513, 2003.

[8] HRD. 2004 hurricane field program plan. Available from the Hurricane Research Division.

[9] A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[10] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[11] E. Kolatch. Clustering algorithms for spatial databases: A survey. Dept. of Computer Science, Univ. of Maryland, 2002.

[12] W. P. Menzel and J. F. Purdom. Introducing GOES-I: The first of a new generation of geostationary operational environmental satellites. *Bulletin of the American Meteorological Society*, 75(5):757–780, 1994.

[13] R. Ng and J. Han. Clarans: A method for clustering objects for spatial data mining. *IEEE Transaction on Knowledge and Data Engineering*, 14(5):1003–1016, 2002.

[14] OFCM. National hurricane operations plan, publication FCM-P12-2004. Available from the Office of the Federal Coordinator for Meteorological Services and Supporting Research.

[15] K. Ooyama. Scale-controlled objective analysis. *Monthly Weather Review*, 115(10):2479–2506, 1987.

[16] K. Ooyama. The cubic-spline transform method: Basic definitions and test in a 1d single domain. *Monthly Weather Review*, 130(10):2392–2415, 2002.

[17] M. D. Powell. The transition of Hurricane Fredric boundary-layer wind field from the open Gulf of Mexico to landfall. *Monthly Weather Review*, 110(12):1912–1932, 1982.

[18] M. D. Powell. Changes in the low-level kinetic and thermodynamic structure of Hurricane Alicia (1983) at land fall. *Monthly Weather Review*, 115(1):75–99, 1987.

[19] M. D. Powell and S. Houston. Hurricane Andrew's landfall in South Florida. Part II: Surface wind fields and potential real-time applications. *Weather and Forecasting*, 11(9):329–349, 1996.

[20] M. D. Powell, S. Houston, and T. A. Reinhold. Hurricane Andrew's landfall in South Florida. Part I: Standardizing measurements for documentation of surface winds fields. *Weather and Forecasting*, 11(9):304–328, 1996.

[21] L. J. Shapiro. The asymmetric boundary layer flow under a translating hurricane. *Journal of the Atmospheric Sciences*, 40(8):1984–1998, 1993.

[22] G. Sheikholeslami, S. Chatterjee, and A. Zhang. Wavecluster: a multi-resolution clustering approach for very large spatial databases. In *Proceedings of the 24th VLDB Conference*, pages 428–439, New York, NY, 1998.

[23] E. W. Uhlhorn and P. G. Black. Verification of remotely sensed sea surface winds in hurricanes. *Journal of Atmospheric and Oceanic Technology*, 20(1):99–116, 2003.

[24] W. Wang, J. Yang, and R. Muntz. Sting: a statistical information grid approach to spatial data mining. In *Proceedings of the 23rd VLDB Conference*, pages 186–195, Athens, Greece, 1997.

[25] X. Xu, M. Ester, H.-P. Kriegel, and J. Sanders. A distribution-based clustering algorithm for mining in large spatial databases. In *Proceedings of the Fourteenth International Conference on Data Engineering*, pages 324–331, Orlando, FL, February 1998. IEEE Computer Society.