APPLICATION OF A NONPARAMETRIC APPROACH TO ANALYZE ΔpCO_2 DATA FROM THE SOUTHERN OCEAN

Wesley B. Pretorius*[1], Sonali Das[2] and Paul J. Mostert[1]

*1 Cell: 0722897595, Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland 7602, 15023133@sun.ac.za

2 Logistics and Quantitative Methods, CSIR Built Environment, PO Box 395, Pretoria 0001

* To whom correspondence should be addressed

APPLICATION OF A NONPARAMETRIC APPROACH TO ANALYZE ΔpCO_2 DATA FROM THE SOUTHERN OCEAN

ABSTRACT

In this paper we discuss the application of a classical nonparametric inference approach to analyse ΔpCO_2 measurements from the Southern Ocean, which is a novel method to analysing data in this area, as well as comparing results with the regular parametric approach. ΔpCO_2 is the difference between atmospheric and ocean partial pressure of CO_2. Oceans are estimated to absorb about 40% of anthropogenic carbon dioxide emissions and can act as both a carbon sink as well as a carbon source. The Southern Ocean, which comprises a large part of world oceans, thus plays a crucial role in the balance of atmospheric CO_2. However, the region south of Africa is largely unanalysed due to data from the region being only very recent. In this paper we analyse *in situ* measurements of ΔpCO_2 data obtained from the Antarctic to Cape Town leg of the SANAE 49 trip during February 02 – 22, 2010. We use a nonparametric approach to understand the behaviour of the distribution of the *in situ* ΔpCO_2 measurements as analysis reveals that the distribution of the data is not unimodal, indicating that traditional parametric methods may not capture its distribution well.

KEYWORDS: Southern Ocean; Carbon cycle; ΔpCO2; Nonparametric; Gaussian kernel.

1 INTRODUCTION

This paper presents an analysis of recent, *in situ* data from the SANAE49 ship leg 6 that consists of the ship‟s journey from the Antarctic to Cape Town during February 02– 22, 2010. Carbon Dioxide (CO_2) is widely regarded as being the gas most responsible for global warming. It is suggested that due to CO_2 emissions by man (anthropogenic CO_2 emissions), the levels of CO_2 in the atmosphere have climbed to more than 30% higher than they were before the industrial revolution. (Barnola, 1999; Keeling & Whorf, 2000) The clearing of forests and the harvesting of wood also reduces carbon-bearing vegetation (i.e. vegetation which extract and retain carbon dioxide to use in the manufacturing of their food), and have been equally

responsible for the increase in the atmospheric carbon dioxide levels (Sarmiento & Gruber, 2002; Houghton, 2001).

Bakker *et al.* (1997) suggest that the increased atmospheric CO_2 levels account for approximately 60% of the emissions made by the burning of fossil fuels. This is due to the retention of CO_2 by the plants and soils, as well as the ocean, referred to as natural carbon sinks (Sermiento & Gruber, 2002). Some sources suggest that the oceanic sinks represent retention of about 17% - 39% of the CO_2 produced by the burning of fossil fuels by humans. (Sarmiento & Sundquist, 1992; Schimel *et al.,* 1995; Siegenthaler & Sarmiento, 1993; Tans *et al.,* 1990) Therefore a model allowing us a deeper understanding of the air-sea carbon flux cycle is very crucial.

The Southern Ocean (particularly the area South of South Africa and North of Antarctica) represents an area of very little research in comparison to the Northern oceans due to difficult sampling conditions, as well as limited times that data is able to be collected. Statistical methods used to analyse oceanic CO_2 data has. Telszewski *et al.* (2009) proposed the usage of self organising neural networks to estimate the pCO_2 distribution in the North Atlantic Ocean. The results offer a viable and accurate model, especially in the summer months, however a major disadvantage of using neural networks is the lack of a simple interpretation which could be communicated to non-statistical professionals. Takhashi *et al.* (2002) indicate that the area between 40°S and 60°S of the Equator (which is the area described between Antarctica and South Africa) represents areas of strong oceanic CO_2 sinks. The partial pressure of Carbon Dioxide (pCO_2) can be described as the pressure of the gas phase CO_2 (above the water) that would occur when no more CO_2 is dissolved by the water (i.e. the reaction is at an equilibrium point). This pCO_2 acts as a proxy for the concentration of CO_2 in the water (or atmosphere when dealing with atmospheric pCO_2) and is explained by a combination of physical, climatological and bio-geochemical factors. Low pCO_2 levels in the water, combined with high wind speeds, increase the CO_2 uptake by these waters. In fact, it is suggested that the ocean south of 50°S, which encompasses only approximately 10% of the global oceanic area, is responsible for about twenty percent of the oceanic uptake of CO_2. (Takahashi *et al.* 2002)

This paper presents new information obtained from *in situ* measurements during 2009/10 in the form of a statistical analysis of the carbon flux in the Southern Ocean between Antarctica and Cape Town. The aim is to develop a model that can capture the distribution of the ΔpCO_2. Section 2 discusses the cleaning procedures used to obtain the data for analysis which we use. Section 3 then builds a nonparametric model on the clean data obtained from section 2. Finally section 4 discusses the results concludes.

2 DATA

The original data obtained from the SANAE49 ship travelling on leg six will henceforth be referred to as SANAE49L6. It consisted of 9215 rows of *in situ* measurements on 27 columns of variables. Measurements started on 12/02/2010 (GPS time 00:04:48) and ended on 22/02/2010 (GPS time 23:55:54), travelling between (70.6245°S, 0.0001°W) and (34.073°S, 17.4585°E). “Spikes” in the data were identified graphically and manually removed from the data set after confirming that they were the result of some form of measurements error. The reduced dataset started on 13/02/2010 (GPS time 18:07:50) and ended on 21/02/2010 (GPS time 18:30:53) travelling between (69.5998°S, 5.9036°W) and (37.0004°S, 12.918°W), and will be referred as SANAE49L6-ver2.

To obtain atmospheric pCO_2 measurements (which were more sparse) to the same scale as the water pCO_2 measurements, Euclidean weighted averaging method was used. By this process, a value for the atmospheric pCO_2 was imputed using the nearest above and below measurement, and a weighting inversely proportional to the distance. The atmospheric and water pCO_2 values were then subtracted from one another to obtain the ΔpCO_2 column, which indicates the air-sea CO_2 flux. Negative values of ΔpCO_2 indicate a source and positive values indicate a sink. The resultant data set, referred to as SANAE49L6-final, contained 6105 rows and ran between the same times and co-ordinates as SANAE49L6-ver2. Table 1 indicates the variables along with a short explanation in SANAE49L6-final that are relevant to our analysis. Figure 1 plots ΔpCO_2 versus latitude co-ordinate of the measurement. This figure suggests an abrupt change in the ΔpCO_2 measurements from positive to negative around 60°S of the equator, highlighting the need for a more in depth knowledge of the carbon flux in this area.

TABLE 1

VARIABLE EXPLANATION FOR SANAE49L6-FINAL

Variable	Explanation
Date	Date of Measurement (mm/dd/yyyy)
gps time	Time of Measurement (hh:mm:ss)
latitude	Latitude Measurement (Negative = South)
longitude	Longitude Measurement (Negative = West)
Salinity	Of the Water
O2(%%%%sat)	Oxygen % Saturation (about right but not calibrated)
O2(ppm)	Oxygen (mg/l) (about right but not calibrated)
pH	Of the Water (Not accurate but diagnostically useful in relative units)
Ch.conc	Chlorophyll: Fluorescence Units (not calibrated)
Intake.Temperature	Outside Sea Surface Temperature
pCO2W(H2OSST)	Water pCO_2 corrected for H2O and SST
MLD	Mixed Layer Depth (Meters)

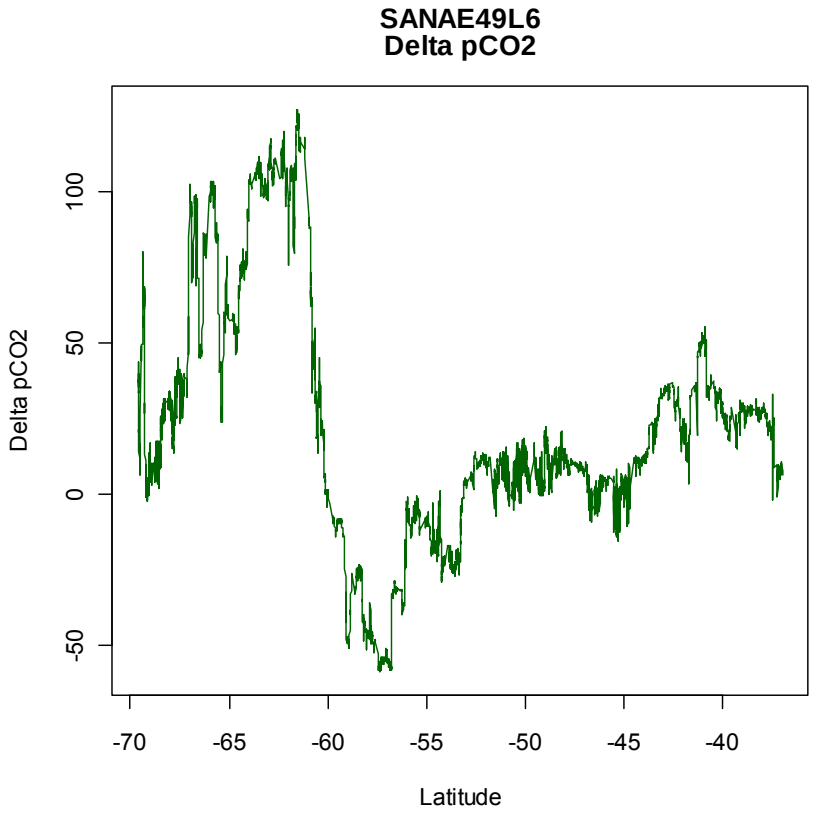

FIGURE 1: LATITUDE PLOT OF DELTA pCO_2

NONPARAMETRIC MODELLING

This section discusses using nonparametric kernel methods in order to estimate the density of the ΔpCO_2 measurements which were obtained from the interpolated

atmospheric pCO_2 measurements discussed in the previous section. The results are compared to parametric density estimation using a Normal (Gaussian) distribution to model the density of the ΔpCO_2.

Nonparametric Kernel density estimation using empirically determined and rule of thumb bandwidths and the Gaussian kernel (Li & Racine, 2007) was used in order to estimate the density function of the ΔpCO_2 measurements. The Gaussian kernel was used with along with the bandwidth definitions that allowed for the most smooth resulting density function. Other kernels resulted in similar, but more volatile densities, indicating a worse, more variable, fit. Three methods for determining the bandwidths were used. Firstly a rule of thumb defined by Silverman (1986) for the Gaussian Kernel which takes a bandwidth of 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth (This resulted in a bandwidth of 3.7). Secondly a variation of this rule of thumb was used suggested by Scott (1992), which uses a factor of 1.06 instead of 0.9 (resulting in a bandwidth of 5.5). Finally biased cross validation was also used to determine an empirically optimum bandwidth (the resulting bandwidth was 4.35). A biased cross-validation method was used since it has a greater reliability than unbiased cross-validation (using the minimization of the sum of square errors), specifically when calculating an optimum smoothing parameter for density estimation (Scott & Terrell, 1987). Nonparametric density estimation is done using the function *density* in R which incorporates all the described methods for determining an optimum bandwidth, while parametric (Normal) density estimation is done by applying a Normal distribution using the mean of the observed ΔpCO_2 measurements as an estimate for the Normal mean and the standard deviation of the observed ΔpCO_2 measurements as an estimate for the Normal standard deviation.

The *density* and *pnorm* functions in R were used to provide estimated density functions of the observed ΔpCO_2 measurements and then the fits of the estimated densities were assessed by comparing the estimated cumulative distribution function to the empirical distribution function of the data. The fits of the estimated distributions are assessed in the following section.

4 DISCUSSION AND CONCLUSION

This section discusses the results of the nonparametric as well as the parametric density estimation in the previous section which are then compared. Finally a conclusion is made as to which of the 2 methods allows for the most accurate estimate of the ΔpCO_2 density. Figure 2 shows the histogram of the ΔpCO_2 measurements indicating the density plot rather than the frequencies. Overlaid on the plot in red is the Normal (Gaussian) density and in purple, green and blue, the nonparametric Gaussian kernel estimated density functions, using the 3 different methods for determining an optimal bandwidth as described above, of the ΔpCO_2 measurements.

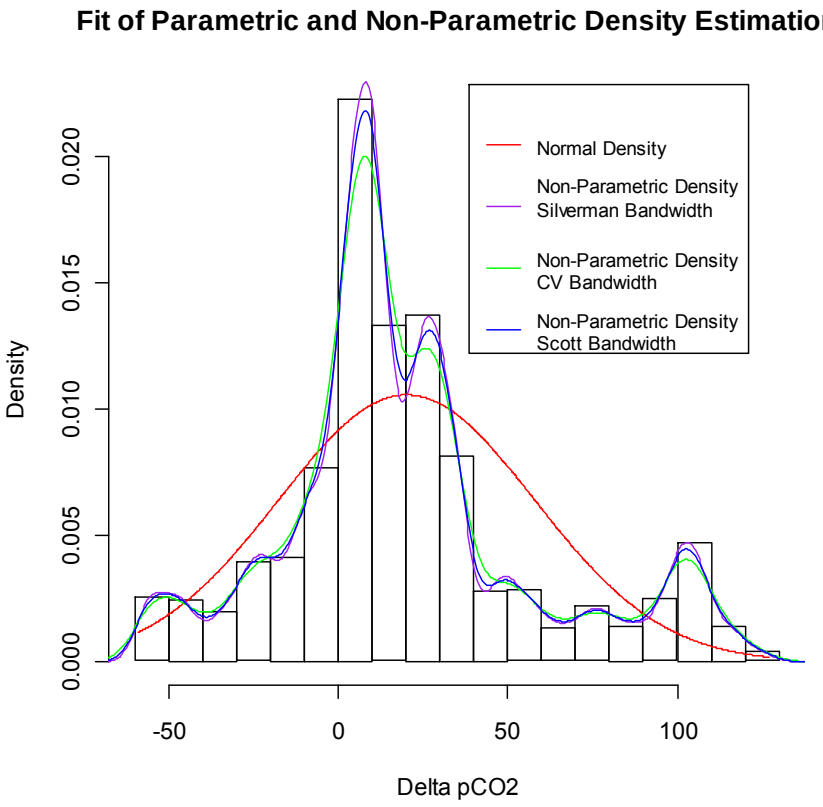

FIGURE 2: DENSITY ESTIMATION COMPARISON

As can be seen, the nonparametric density estimations seem to capture the functional form of the data density much better than the parametric Normal distribution. The normal plot does not capture the high density of values of ΔpCO_2 that are slightly larger than 0, as well as over estimates the densities for values of ΔpCO_2 that are above 50 and those between 0 and -50 units. Another feature not identified by the Normal density estimation is the increased density for values slightly larger than 100.

The nonparametric density estimation graphs, however, capture all these features and seem to follow the histogram of the measured values well.

Figure 3 below provides a graphical goodness of fit test for the nonparametric density estimation process. The plot compares the cumulative density function (CDF) of the estimated densities for the parametric and nonparametric cases to the empirical CDF of the ΔpCO_2 measurements. It is clear in the plot that the CDF's for the nonparametric method are almost exactly the same as the empirical CDF, while the parametric (Gaussian) CDF seems to not capture the form of the data, specifically in the 3 areas discussed earlier as areas which most display the parametric density's lack of fit. This indicates that the nonparametric density estimates represent good fits to the ΔpCO_2 measurements, while the parametric density estimate does not capture the form of the data well enough and therefore constitutes a bad fit.

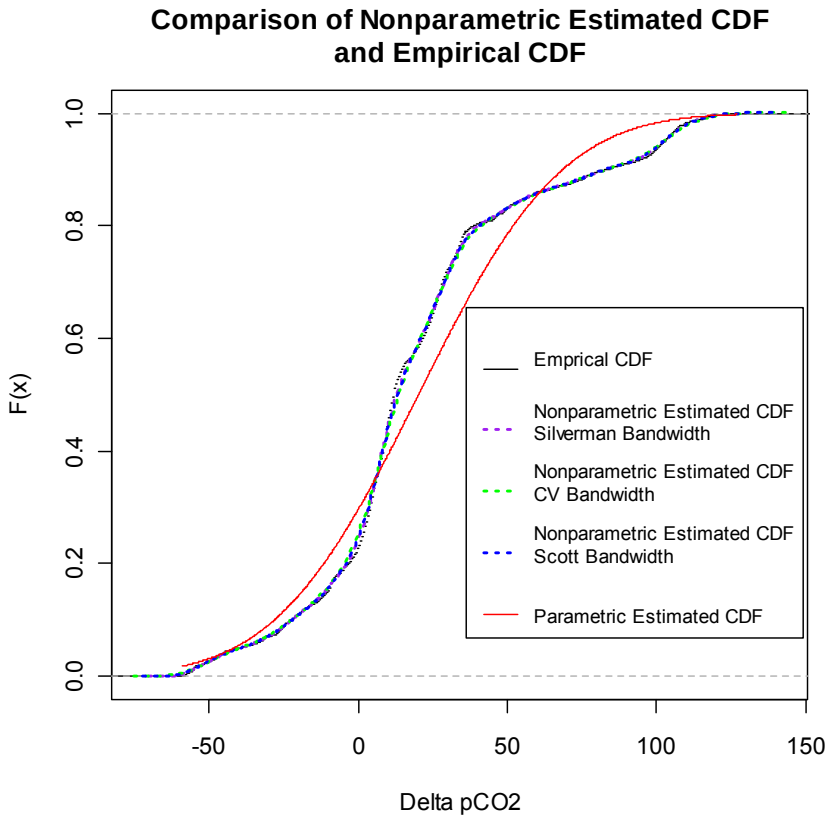

FIGURE 3: COMPARISON OF NONPARAMETRIC AND EMPIRICAL CDF'S

The Kolmogorov-Smirnoff goodness-of-fit test was also conducted on the Normal (Gaussian) density estimation in order to provide a second, more quantifiable,

assessment of the lack of fit for the parametric density. This test was carried out using the *ks.test* function in the *stats* package of R. The test applies to the null hypothesis that the parametric (Gaussian) density is a good fit for the data versus the two-sided alternative that it does not fit. A summary of the results of the test is contained in table 2 below.

TABLE 2

SUMMARY OF KOLMOGOROV-SMIRNOV GOODNESS-OF-FIT TEST

Test Statistic (D)	P-Value	Decision
0.1292	0	Reject the Null Hypothesis at any Significance Level

The result of the test indicates that the null hypothesis that the parametric (Gaussian) density is a good fit for the data will be rejected at any significance level and therefore this provides additional evidence to support the claim that the parametric density estimation does not capture the density of the data well enough.

The analyses presented in this paper corroborates the motivation that nonparametric density estimation provides a much better fit of the ΔpCO_2 measurements than the Normal (Gaussian) distribution.The parametric density estimation approach seems to provide an inadequate fit for the data, as indicated by the Kolmogorov-Smirnov hypothesis test, and therefore it is suggested that nonparametric methods be used in modelling this univariate problem. This is due to the fact that the graphical goodness-of-fit tests indicate that the nonparametric density estimation provides a better fit to the data. Nonparametric kernel methods also provide a model which is easier to interpret to persons who may not have a statistical background than black box methods such as a neural network. This analysis, therefore provides an alternative modelling approach to that suggested by Telszewski *et al.* (2009). Had a parametric approach been applied, it is clear that the density of the ΔpCO_2 would be underestimated for mid-valued ΔpCO_2 values and over estimated for higher and lower values of ΔpCO_2 and then again underestimated in the boundary values of ΔpCO_2.

Possible future research directions to analysing this data are to incorporate covariate information such as temperature, salinity, pH and oxygen saturation, chlorophyll concentration and the measure of mixed layer depth. This would result in a

regression type setup, which could be subjected to both parametric and nonparametric modelling in order to determine the model that provides the „best" fit (closer fitting without over-fitting). Bayesian nonparametrics, such as process priors, is another approach which could be used in order to model the data. This would not only allow for more stochastic features in the modelling, but also be a more generalized model for data which may come from different regions. Finally, the use of a mixture distribution seems also to be a useful alternative and may provide more easily interpretable results. A further direction will be to analyse the sudden change in ΔpCO2 measurements from region of sink to a region of source around 60°S of the equator (Figure 1).

5 ACKOWLEDGEMENTS

We thank Prof. Biman Chakraborty of the University of Birmingham for his help with R codes and related discussions. We also acknowledge valuable discussion with domain experts Dr. Nicolas Fauchereau and Dr. Pedro Monteiro, both of the CSIR, Stellenbosch.

REFERENCES

Bakker, D.C.E., Baar, H.J.W. & de Bathmann, U.V. 1997. Changes of carbon dioxide in surface waters during spring in the Southern Ocean. *Deep-Sea Research Part II,* 44:91–128.

Barnola, J.M. 1999. Status of the atmospheric CO_2 reconstruction from ice core analyses. *Tellus,* 51B:151-155

Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X., Maskell, K. & Johnson, C.A. 2001. Climate Change 2001: The Scientific Basis. Cambridge University Press, Eds. 2001. New York, [Online] Available from: *http://www.ipcc.ch*

Keeling, C.D. & Whorf, T.P. 2000. Atmospheric CO2 records from sited in the SIO air sampling network. In Keeling, C.D. & Whorf, T.P. Eds. 2000. Trends: A Compendium of Data on Global Change. Oak Ridge National Laboratory, Oak Ridge.

Li, Q. & Racine, J.S. 2007. Nonparametric Econometrics: Theory and Practice, Princeton University Press, 57-115

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-00051-07-0, [Online] Available from: http://www.R-project.org/.

Sarmiento, J. L. & Sundquist, E. T. 1992. Revised budget for the oceanic uptake of anthropogenic carbon dioxide. *Nature*, 356:589-593.

Sarmiento, J. L., and N. Gruber 2002, Sinks for anthropogenic carbon. *Physics Today*, *55*(8):30-36.

Schimel D., Enting, I. G., Heimann, M., Wigley, T. M. L., Raynaud, D., Alves, D. & Siegenthaler, U. 1995. CO2 and the carbon cycle. In Houghton, J.T., Meira Filho, L.G., Bruce, J., Lee, H., Callander, B.A., Haites, E., Harris, N. & Maskell, K. Eds. 1995. *Climate change 1994. Radiative forcing of climate change and an evaluation of the ZPCC IS92 emission scenarios,* 339, Cambridge University Press, Cambridge, Intergovernmental Panel on Climate Change, 35-71.

Scott, D.W. 1992. Multivariate Density Estimation. Theory, Practice and Visualization. New York:Wiley.

Scott, D.W. & Terell, G.R. 1987. Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association 82(400):1131-1146.

Siegenthaler, U. & Sarmiento, J. L. 1993. Atmospheric carbon dioxide and the ocean. *Nature*, 365:119-125.

Silverman, B. W., 1986. Density Estimation. London: Chapman and Hall.

Takahashi, T., Sutherland, S., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N., Wanninkhof, R.H., Feely, R. & Sabine. C. 2002. Global sea-air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature. *Deep Sea Research Part II: Topical Studies in Oceanography*, 49(9-10):1601-1622.

Tans, P.P., Fung, I. Y. & Takahashi, T. 1990 Observational constraints on the global atmospheric CO_2-budget. *Science,* 247:1431-1438.

Telszewski, M., Chazottes, A., Schuster, U., Watson, A.J., Moulin, C., Bakker, D.C.E., Gonzalez-Davila, M., Johannessen, T., Kortzinger, A., Luger, H., Olsen, A., Omar, A., Padin, X.A., Rios, A., Steinhoff, T., Santana-Casiano, M., Wallace, D.W.R. & Wanninkhof, R. 2009. Estimating the monthly pCO2

distribution in the North Atlantic using a self-organizing neural network. Biogeosciences Discussion, 6:3373-3414.