Article:
Rueda, C, Fernández, MA, Barragan, S et al. (2016) Circular piecewise regression with applications to cell-cycle data. Biometrics, 72 (4). pp. 1266-1274. ISSN: 0006-341X

https://doi.org/10.1111/biom.12512

Biometrics 64, 1–20 DOI: 10.1111/j.1541-0420.2005.00454.x
February 2015

Circular Piecewise Regression with an Application to Cell-cycle Biology

Cristina Rueda[1,†,*], Miguel A. Fernández[1,†,], Sandra Barragán[1,***],**

Kanti V. Mardia[2,**] and Shyamal D. Peddada[3,*****]**

[1] Departamento de Estadística e I.O., Universidad de Valladolid, 47011 Valladolid, Spain

[2] Department of Statistics, University of Oxford, Oxford, UK,

and Department of Statistics, University of Leeds, Leeds, UK

[3] Biostatistics Branch, NIEHS (NIH), Research Triangle Park, NC, USA

[†] These authors contributed equally to the paper

email: crueda@eio.uva.es

**email:* miguelaf@eio.uva.es

***email:* sandraba@eio.uva.es

****email:* K.V.Mardia@leeds.ac.uk

*****email:* peddada@niehs.nih.gov

Summary: Applications of circular regression models appear in many different fields such as evolutionary psychology, motor behavior, biology, and, in particular, in the analysis of gene expressions in oscillatory systems. Specifically, for the gene expression problem, we need to model the relation among peak expressions of cell-cycle genes in two species with different cell phase lengths. This challenging problem reduces to the problem of constructing a piecewise circular regression model and, with this objective in mind, we propose a flexible circular regression model which allows different parameter values depending on sectors along the circle. We give a detailed interpretation of the parameters in the model and provide maximum likelihood estimators. We also provide a model selection procedure based on the concept of generalized degrees of freedom. The model is then applied to the analysis of two different cell-cycle data sets and through these examples we highlight the power of our new methodology.

Key words: Circular data; Circular–circular regression; Change points; Gene expression; Generalized AIC; Von Mises distribution.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Analysis of circular data has a long history with well-developed theory and methodology documented in several books (see, for example, Fisher (1993); Mardia and Jupp (2000)). It is well known that circular data analysis presents many challenges due to the non-Euclidean nature of the circle. All basic concepts have to be redefined as even the usual arithmetic mean is not meaningful in this case, since, for example, the mean value of 1 and 359 degrees cannot be 180 degrees (which is in the opposite side of the circle) but should be 0 degrees which is the circular mean. Until recently much of the literature was developed for describing circular models and drawing inferences on individual angular parameters, such as comparing the mean directions of two or more populations. In recent years, the circular-circular regression problem (i.e. when both regressor and response are circular variables) has received some attention (see, for example, Downs and Mardia (2002), Kato et al. (2008), Kato and Jones (2010) or Polsen and Taylor (2015)). In order to perform circular regression a link between the two variables is needed in order to get meaningful results since there is no concept of scaling/slope on this manifold. With this in mind, it is not surprising that there is not a unique approach to circular regression unlike for linear regression. The most natural model using a tangent link function and based on Möbius transformations was introduced by Downs and Mardia (2002) under the assumption that the angular variables were distributed according to the von Mises distribution; the von Mises distribution is the standard analogue on the circle of the univariate normal distribution. Our piecewise circular regression model introduced in this paper is based on this model. This is a very flexible as required by our motivating applications. In particular, the proposed model is able to deal with the additional complication of the continuity restrictions, computational burden introduced by the piecewise character of the model, and some specific requirements of the our applications.

This paper is motivated by a problem encountered in cell biology where researchers are

interested in correlating angular data from two or more sources (e.g. experiments or species). We now give insight into the data. There are four major phases of distinct biological functions through which a cell-cycle gene (among eukaryotic cells) goes during cell division, namely the four phases are G1, S, G2 and M. These genes participating in the cell-cycle tend to have a periodic pattern of expression over time. Consequently, the time to peak expression (known as the phase angle) of such genes can be mapped onto a unit circle. Further the time boundaries of each phase are mapped in to sectors which for our application in Section 3.2 are $(2.10, 2.80, 4.00, 5.75)$. These sector boundaries are important for this paper and will be denoted by θ_i^*, $i = 1, 2, ..., 4$ (Section 2.1). Figure 1 shows the four gene phases and the sectors so, for example, a gene in G1 takes a value in the sector $(2.10, 2.80)$. From our data set (Application in Section 3.2), the figure shows phase angles (data points) of the four *S. cerevisiae* genes: RFA1, HHT1, FHK1, and DBF2. The figure also shows the cell phase length of each cycle $(0.70, 1.20, 1.75, 2.63)$ and their relative percentages $(11\%, 19\%, 28\%, 42\%)$.

[Figure 1 about here.]

To illustrate the methodology we use cell-cycle data available from the cyclebase data base www.cyclebase.org (Santos et al. (2015)). This database contains data obtained from 20 different experiments conducted in different laboratories on budding yeast (*S. cerevisiae*) and fission yeast (*S. pombe*).

Cell-biologists are often interested in drawing inferences regarding the phase angle of cell-cycle genes since they are considered to be associated with the gene's biological function. Often two types of inferences are of interest. Using the data obtained from a single experiment on a given species, one may be interested in estimating the phase angle of a set of cell-cycle genes in that species when the relative order of peak expression is known *a priori* (Rueda et al. (2009)). Another question of interest is to detect whether the order of the phase angles of a set of cell-cycle genes is consistent across multiple experiments on the same species (Liu

et al. (2004)), or more broadly if the order of the phase angles of a set of cell-cycle genes is the same across multiple species (Fernández et al. (2012)).

For our applications, we need a model with the following features. 1. Monotonicity. In the cell-cycle, the function relating the peak expressions has to be increasing as a decrease in the function would mean that the cycle is going backwards, which is biologically not sensible. 2. "Synchronicity" (defined with more detail in Section 2.3). As we are relating the data coming from a single cycle in the response variable to those coming from a single cycle in the regressor variable, the response has to run one cycle when the regressor variable runs through one cycle. We show in Section 2.3 how these conditions can be incorporated in our model which will not be easy in the non-parametric models such as of Di Marzio et al. (2013).

While, as demonstrated in Liu et al. (2004), the regression model proposed in Downs and Mardia (2002) is likely to perform well when the cell phase lengths are the same across all species, it may be too rigid when the cell phase length in each of the four phases is not the same across different species. For this reason, in Section 2 we introduce a flexible piecewise regression model that can be useful for drawing inferences when the cell phase lengths vary across species.

Piecewise regression, although not defined for manifolds until now, has been well studied in the Euclidean setting (see for example Seber and Wild (1989)). To highlight some challenges in circular piecewise regression, we consider the simplest linear case. Namely, the case of a single change point with no error

$$y = a_1 + b_1 x, x \leqslant c; \quad y = a_2 + b_2 x, x \geqslant c$$

with the continuity constraint

$$a_1 + b_1 c = a_2 + b_2 c. \tag{1}$$

We note that if x and y are are angular variables, a single change point c has no meaning because a single point does not define two sectors in the closed circumference, so there should

be at least two change points; the two sector boundaries could consists of, say, day and night. (We note that our applications in Section 3 show four change points). Furthermore, in the linear case, this problem for computational purpose can be reparametrized as

$$y = A + Bx + C(x - D)SGN(x - D) \tag{2}$$

where now

$$c = D,\ a_1 = A + CD,\ a_2 = A - CD,\ b_1 = B - C,\ b_2 = B + C$$

so the constraint (1) is included in (2). As noted later in the paper, such a simplification is not available for the circular case. Of course, when the noise is added the inference problems become more intricate as we will see in Section 2.

The methodological contributions of this paper are provided in Section 2 where we develop the piecewise circular regression model and interpret the parameters of the model. We then describe the estimation of these parameters. In that section we also describe a model selection procedure based on the Generalized Akaike Information Criterion.

In Section 3 we illustrate our methodology by applying it to a cell-cycle gene expression that motivated this study. In the first example, both data sets are on the same species but obtained from different laboratories. In the second example, the data are obtained from two different species of yeast, namely, fission yeast and budding yeast. Finally, in Section 4, we discuss a variety of other biological applications, and explain the flexibility of our model for different applications. We also point out various extensions of our work.

2. The circular piecewise regression model

In Section 2.1, we define the piecewise circular regression model that extends Downs and Mardia (2002) while allowing for flexible relationships between regressor and the response variable in different sectors of the circle. In the subsequent sections, Section 2.2–Section 2.3, we interpret the parameters of the model and derive the maximum likelihood estimators. We

demonstrate that our model is flexible to allow for monotonicity by imposing a restriction on the parameters. In Section 2.4 we define a model selection criterion based on generalized degrees of freedom by Ye (1998).

2.1 *The Model*

Consider a circular response variable ψ and a circular regressor variable θ. We denote by k the number of different pieces or sectors in the unit circle and as θ_i^*, $i = 1, 2, ..., k$ the sector boundaries (or change/break points in the linear piecewise regression model in the line) which are assumed to be known. Note that $k > 1$ as we need two change points to define two sectors on the circle. We denote as $\mathbf{\Theta}$ the vector of values for the regressor variable with components θ_{ij}; $i = 1, 2, ..., k$ and $j = 1, 2, ..., n_i$, where the first index is the sector the observation belongs to, so that $\theta_i^* < \theta_{ij} \leqslant \theta_{i+1}^*$ for $1 \leqslant i \leqslant k-1$ and the index i takes value k when $\theta_k^* < \theta_{ij} \leqslant 2\pi$ or $0 \leqslant \theta_{ij} \leqslant \theta_1^*$, the second index j is the number of the observation in the corresponding sector, n_i is the number of observations in sector i and $N = \sum_{i=1}^{k} n_i$ is the total number of observations. Accordingly we denote as $\mathbf{\Psi}$ the vector of observed values and as ψ_{ij} the corresponding components of this vector. We further assume that ψ_{ij} given θ_{ij} comes from independent von Mises distributions $M(\mu_{ij}, \kappa)$ with density function $f(\theta_{ij}, \mu_{ij}, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta_{ij} - \mu_{ij})}$, where I_0 denotes the modified Bessel function of the first kind and order 0.

We now describe the circular–circular (c-c) regression model of Downs and Mardia (2002) where we take the circular response variable ψ, and the circular regressor variable θ, but in this case there are no boundaries for θ. Further, ψ given θ comes from a von Mises distribution with $M(\mu, \kappa)$ and

$$\tan\frac{1}{2}(\mu - \beta) = \omega \tan\frac{1}{2}(\theta - \alpha),$$

where α and β are angular location parameters and ω is a slope parameter which is restricted to the closed interval $[-1, 1]$ by adjusting α and β appropriately. Next, we propose our

piecewise circular regression model with

$$\tan\frac{1}{2}(\mu_{ij}-\mu)=\omega_i\tan\frac{1}{2}(\theta_{ij}-\nu_i),\;\; j=1,\ldots,n_i; i=1,\ldots,k, \tag{3}$$

where, to ensure continuity, we take

$$\omega_i\tan\frac{1}{2}(\theta_i^*-\nu_i)=\omega_{i-1}\tan\frac{1}{2}(\theta_i^*-\nu_{i-1}),\;\; i=1,\ldots,k, \tag{4}$$

and $\nu_0=\nu_k$, $\omega_0=\omega_k$. This model maintains the functional relationship of the Downs and Mardia model but allows a different parameters in each of the sectors while imposing continuity on the global function. Equivalently, our model (3) can be rewritten as

$$\mu_{ij}=\mu+2\arctan\left(\omega_i\tan\frac{1}{2}(\theta_{ij}-\nu_i)\right), \tag{5}$$

for $j=1,\ldots,n_i$ and $i=1,\ldots,k$.

When $k=2$ there is a simplification of the model given in Web Appendix A. Note also that we cannot restrict all the ω_i values to the interval $[-1,1]$ for the piecewise model. In the c-c model the ω parameter can be restricted to that interval as an equivalent model can be obtained considering $\mu'=\mu-\pi$, $\omega'=1/\omega$ and $\nu'=\nu-\pi$. However, in the piecewise model it is not possible to make a transformation to ensure that $\omega_i\in[-1,1]$ simultaneously for all $i=1,\ldots,k$ as that would require more than one value for the single μ parameter. In spite of this, from equation (3) it is clear that, as in the c-c model, the model with parameters $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_k)$, $\boldsymbol{\nu}=(\nu_1,\ldots,\nu_k)$ and μ is completely equivalent to that with parameters $\boldsymbol{\omega'}=(1/\omega_1,\ldots,1/\omega_k)$, $\boldsymbol{\nu'}=(\nu_1-\pi,\ldots,\nu_k-\pi)$ and $\mu'=\mu-\pi$.

2.2 *Interpretation of the model parameters*

Since the meaning of the parameters in our model is not straightforward (as for example in the normal linear regression model), we give a detailed interpretation of each of them. Parameter μ can be easily interpreted as a global location parameter quantifying the rotation of the response that allows effective alignment with the regressor variable. Since there are different sectors, different rotation parameters in each of the sectors are also needed for an

appropriate alignment between the regressor variable and the response; this is accounted for by the ν_i parameters.

The ω_i are slope parameters for the θ_{ij} observations in the sector $(\theta_i^*, \theta_{i+1}^*]$ but its interpretation is not that easy. If we leave aside the μ parameter, notice that each set of (ω_i, ν_i) parameters corresponds to one of the curves appearing in (5). From each of these k curves, the model only uses a sector of length $\theta_{i+1}^* - \theta_i^*$ for sector i when $i = 1, \ldots, k-1$ and of length $2\pi - \theta_k^* + \theta_1^*$ for sector k. The sector used is that from the interval $\left(\theta_i^* - \nu_i, \theta_{i+1}^* - \nu_i\right)$ when $i = 1, \ldots, k-1$ and $(\theta_k^* - \nu_k, \theta_1^* + 2\pi - \nu_k)$ for $i = k$. As the model is not linear, a higher ω_i parameter (we will assume $\omega_i > 1$ in this paragraph) does not always mean a steeper curve in the model. A high ω_i parameter corresponds to a steep curve in the central part of the $[0, 2\pi]$ interval and to a flat curve at both extremes of that interval. Therefore, as only a part of the curve corresponding to the aforementioned intervals is used, the steepness in a sector does not only depend on ω_i but also on ν_i and on θ_i^* and θ_{i+1}^*.

To illustrate these points we consider the following example coming from the data analyzed in Section 3.2. That is, we assume a model with the known parameters (estimated in Section 3.2) as given in Table 2 for which the sector boundaries considered are $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (2.10, 2.80, 4.00, 5.75)$. From these parameters it may appear that the slope on the first sector $(2.10, 2.80]$ ($\omega_1 = 1.658$) should be lower than in the fourth one $(5.75, 2.10]$ ($\omega_4 = 6.517$) while it is clear from the model graph at Figure 2 (bottom) that it is not so. From the same Figure, when compared with the third sector $(4, 5.75]$, it is also apparent that the slopes in the sectors are not constant. The ω_3 parameter of this third sector (65.711) is the highest of the four but there are some parts of the sector where the curve is not steep.

To give further detail on the relationship between the ν_i and the ω_i parameters, consider Figure 2 (top left); it gives the four curves from which the model is built. The blue curve corresponds to $\omega_1 = 1.658$, the red one to $\omega_2 = 0.066$, the grey one to $\omega_3 = 65.711$ and the

green one to $\omega_4 = 6.517$. The ν_i parameters and θ_i^* values determine which part of each curve is used in the model. Depending on ν_i and θ_i^*, the same ω value can translate into a more or less steep curve.

[Figure 2 about here.]

Figure 2 (top right) shows how these curves are combined in Figure 2 (top left) using the ν_i parameters, finally leading to the model graph as shown in the bottom of the Figure 2 (bottom). The solution appearing in Figure 2 (bottom) follows the thin blue line from 2.10 (where it intersects the green line for the second time) to 2.80 (where the thin blue line intersects the red line), then the red line from 2.80 to 4.00 (where the red line intersects the grey line), next the grey line from 4.00 to 5.75 (where the grey line intersects the green line) and finally the green line from 5.75 to 2π and from 0 to 2.1.

2.3 *Estimation*

We now describe the derivation of maximum likelihood estimators by assuming that the response components are independently distributed according the von-Mises distribution $M(\mu_{ij}, \kappa)$, where the log-likelihood is given by

$$\max_{\kappa,\mu,\omega,\nu} \left[-N \ln I_0(\kappa) + \kappa \sum_{i=1}^{k} \sum_{j=1}^{n_i} \cos\left(\psi_{ij} - \mu - 2 \arctan\left(\omega_i \tan \frac{1}{2}(\theta_{ij} - \nu_i) \right) \right) \right]. \quad (6)$$

This expression has to be maximized under the continuity condition (4) which leads to the following constraints on the parameters for the existence of a non-trivial continuous solution.

$$\begin{aligned} \tan\left(\frac{\theta_1^* - \nu_k}{2}\right) \prod_{i=1}^{k-1} \tan\left(\frac{\theta_{i+1}^* - \nu_i}{2}\right) &= \prod_{i=1}^{k} \tan\left(\frac{\theta_i^* - \nu_i}{2}\right) \\ \omega_i &= \frac{\tan\left(\frac{\theta_{i+1}^* - \nu_{i+1}}{2}\right)}{\tan\left(\frac{\theta_{i+1}^* - \nu_i}{2}\right)} \omega_{i+1} \text{ for } i = 1, \ldots, k-1. \end{aligned} \quad (7)$$

The above constraints are general for our piecewise circular regression model. However,

recall that two additional conditions have to be imposed for the cell-cycle application. One is the condition for monotonicity and another is of "synchronicity" which we now describe. The monotonicity condition ensures that the solution is an increasing function which leads to the following constraints:

$$\omega_i \geqslant 0 \text{ for } i = 1, \ldots, k. \tag{8}$$

By "synchronicity" condition, we mean that the response runs through only one cycle as the regressor variable runs one cycle. For this purpose, we impose the constraints that the solution only crosses the 0 barrier once. Let

$$z_i = \nu_i + 2\arctan\left(\frac{1}{\omega_i}\tan\left(\frac{-\mu}{2}\right)\right) \text{ for } i = 1, \ldots, k.$$

The z_i value is the possible zero of i^{th} piece of the function. It will be a zero of the global function if this value belongs to the appropriate interval. Thus, the constraints under the synchronicity condition can be written as

$$\sharp\left\{z_i : z_i \in \left(\theta_i^*, \theta_{i+1}^*\right]\right\} = 1, \tag{9}$$

with $\theta_{k+1}^* = \theta_1^*$. Hence, we need to optimize the log-likelihood (6) with the additional constraints given by (8) and (9).

Implementation. In order to compute the maximum likelihood estimates of our model, we rewrite the model as the following piecewise circular-linear model

$$\tan\frac{1}{2}(\mu_{ij} - \mu) = \omega_1 X_1 + \ldots \omega_k X_k,$$

where the regressor variables X_i are defined as $X_i = \tan\left(\frac{\theta_{ij}-\nu_i}{2}\right) I_{(\theta_i^* < \theta_{ij} \leqslant \theta_{i+1}^*]}$. Now, we can compute the maximum likelihood estimates of $\{\kappa, \mu, \omega_1, \nu_1, \ldots, \omega_k, \nu_k\}$ using the theory developed by Fisher and Lee (1992) for circular-linear models, and the R package of Agostinelli and Lund (2011). The optimization has to be performed under the constraints given by (7), (8) and (9). This makes the problem more complex. To solve it, we chose the solution maximizing (6) after repeating the following procedure. We first chose values for

the ν_i parameters verifying the restrictions (7). Then, for these values, we optimized the log-likelihood function (6) in ω_k (as according to (7) the rest of the ω_i values can be expressed as a function of ω_k and the ν_i parameters) taking into account the restrictions (8) and (9). Web Appendix A includes the results of a numerical study showing the performance of this procedure when conditions such as the number of points per sector, length of the sectors or variability parameter κ change.

2.4 *Model selection*

In applications of our piecewise model, we need a procedure to assess the performance of our model relative to the performance of the Down and Mardia model for fitting the angular data. The question of model selection in the Euclidean space setting is well discussed in the literature going back to the seminal paper by Akaike (1973). However, to the best of our knowledge, the model selection problems have not been well addressed for circular models.

For the Euclidean space data a simple strategy used for measuring how well a model fits the data is to compute model residuals. Analogously, in the case of circular data, one may use the circular residuals defined as $e_{ij} = 1 - \cos\left(\psi_{ij} - \left(\hat{\mu} + 2\arctan\left(\hat{\omega}_i \tan\frac{1}{2}(\theta_{ij} - \hat{\nu}_i)\right)\right)\right)$, and define a circular distance criterion (CDC) as $CDC = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}$.

Although, for a pair of competing models, the difference between the two CDC values may be insightful, it is not an ideal measure because a complex model, such as the proposed piecewise model, will always have a lower CDC than a simpler one, such as the c-c model. Alternatively, suppose $l(M)$ denotes the log-likelihood corresponding to model M. Then one may consider the Akaike Information Criterion (AIC), defined as $AIC(M) = 2\ln(l(M)) - 2D$, which corrects for the number of free parameters in the model M using the penalization factor D. A model is selected which gives the largest value of $AIC(M)$. Akaike (1973) suggested $D = p$ which represents the number of parameters in the model. The choice of D is not always straightforward for complex models such as lasso (Tibshirani and Taylor (2012)),

the mixed effects models (Muller et al. (2013)), semiparametric additive monotone models (Rueda (2013)) or in the present situation where the parameters from a circular model are subject to complicated constraints. More precisely, in our piecewise model, we have $2k + 1$ parameters and k restrictions (7) which reduce the number of *free parameters*, to $1 + k$. However, it is not clear how to incorporate restrictions (8) and (9) to get an exact count of the number of free parameters D.

For the case of piecewise linear regression some authors have also used modified AIC or Bayesian Information Criteria (BIC) for selecting the best model (Muggeo and Adelfio (2011); Painting and Holwell (2013)), although these proposals obviously do not take into account the restriction (9) or the manifold we are considering in this paper.

Suppose $\mathbf{Y} = (y_1, ..., y_N)'$ denotes the observed data and its estimated mean vector (which is a function of $\mathbf{Y}$) is given by $\hat{\mu}(\mathbf{Y}) = (\hat{\mu}_1(\mathbf{Y}), ..., \hat{\mu}(\mathbf{Y})')$. Then Ye (1998) defined the generalized degrees of freedom GDF as $\sum_{i=1}^{N} \frac{\partial}{\partial \mu_i} \hat{\mu}_i(\mathbf{Y})$. According to Ye (1998), the GDF is the "sum of the sensitivity of each fitted value to perturbations in the corresponding observed value". This concept was later extended to other models by Zhang et al. (2012). Not only is GDF applicable to complex modeling procedures but is also easy to evaluate as noted in Web Appendix B. In the case of normal linear regression model with p parameters, the $GDF = p$. Hence, motivated by Ye (1998), in this article we use the Generalized Information Akaike Criterion given by,

$$GAIC = AIC(M) = 2\ln(l(M)) - 2\widehat{GDF}$$

for evaluating and comparing models, with the model with the largest $GAIC$ value being the preferred one. Details regarding the derivation of $\widehat{GDF}$ are provided in Web Appendix B. The numerical study in Web Appendix A also contains results showing how this criterion works for distinguishing between the situations where the Downs and Mardia (2002) model works well or the piecewise model is needed.

3. Applications

When a study or experiment is conducted/repeated by multiple labs then it is common to ask how reproducible the data and results are. With the advent of microarray technology, during the past decade multiple labs conducted cell-cycle experiments to compare phases of cell-cycle genes in the genome of various species. However, researchers have been concerned about the reproducibility of results across labs even within the same species. If the phase angles within the same species, obtained from different labs or experiments correlate poorly with each other, then it will be difficult to compare phase angles of cell-cycle genes across multiple species. There is a need for a methodology to assess relations among phase angles between a pair of experiments. However, with the exception of the geometric approach of Liu et al. (2004), to the best of our knowledge there does not seem to exist a formal statistical procedure to make such diagnostics for angular data. Often biologists use visual displays such as heatmaps when assessing similarities between experiments or studies. Such graphical tools ignore variability in the data and hence are not very satisfactory.

Using some recently published cell-cycle data, in the following we demonstrate that the methodology developed in this paper can be used to assess correlations between a pair of experiments. We consider two examples. In the first example we apply our methodology on a pair of experiments conducted in two different labs on the same species of yeast, namely, *S. cerevisiae*. In the second example we apply the methodology on data from two different species of yeast, namely, *S. cerevisiae* and *S. pombe*, conducted in two different labs.

3.1 *Within species between labs correlation of phase angles of cell-cycle genes*

In this example we considered phase angle estimates of 32 *S. cerevisiae* cell-cycle genes obtained from the Spellman cdc experiment (Spellman et al., 1998) and from Pramilla38 experiment (Pramila et al., 2006). The phase angle data obtained from the Spellman cdc experiment were taken to be regressors (θ) and those from Pramilla38 experiment were taken

to be the response variables (ψ). Using the information available in the cyclebase database, we placed the change points at $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (1.50, 2.80, 4.00, 5.75)$. The estimated phase angles, for the 32 cell-cycle genes, the sector they are placed in and a scatterplot of these data can be found in Web Table 3 and in Web Figure 1 respectively.

Using the proposed piecewise circular regression model we obtain the results summarized in Table 1. The fitted model is plotted in Figure 3 while the residuals are in Web Figure 2. For comparison purposes, in each table and graph we also provide results using the c-c model.

[Table 1 about here.]

[Figure 3 about here.]

The value of the CDC measure for the piecewise model is 0.138 while for the c-c model it is 0.148. The CDC has been reduced by 6.76%. To evaluate further whether this improvement is big enough we consider the selection diagnostics we defined in Section 2.4. The GDF values appearing in Table 3 are the mean values obtained by averaging the results of several runs of the GDF algorithm with different values of the tuning parameter τ. The value of $GAIC$ for the piecewise model is 55.474 while for the c-c model it is 55.977, so that there is no evidence of significant improvement due to the piecewise model.

If the role of the variables is reversed, i.e. the Pramilla38 experiment is taken as regressor and the Spellman cdc experiment is taken as response, and the appropriate change points are considered we get the the same result, namely, the $GAIC$ of the piecewise model is lower than that of the c-c model.

As the piecewise model does not yield a significant improvement in this case over the c-c model, we can infer that there is no need for different functional relationships in the different cell phases. In this sense we may also say that there is congruence between these two experiments performed by different laboratories.

3.2 *Between species and between labs correlation of phase angles of cell-cycle genes*

In this example we considered phase angle estimates of 32 cell-cycle genes obtained from two different species *S.cerevisiae* and *S. pombe*. Furthermore, the data were obtained from two different labs. We used the phase angle estimates from the Spellman cdc (Spellman et al. (1998)) on *S. cerevisiae* as the regressors (θ) and those from the Oliva elut2 experiment (Oliva et al. (2005)) on *S. pombe* as the response variables (ψ). For this case, we placed the change points at $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (2.10, 2.80, 4.00, 5.75)$. There is only a small change in the first change point with respect to the previous example that is also congruent with the information available in the cyclebase database. The estimated phase angles for the 32 cell-cycle genes, the sector they are placed in and a scatterplot of the data can be found in Web Table 4 and in Web Figure 3 respectively.

Using the proposed piecewise circular regression model we obtain the results summarized in Table 2. The fitted model is plotted in Figure 4 and the residuals are plotted in Web Figure 4. As in the previous example, for comparison purposes, in each table and graph we also provide results using the c-c model by Downs and Mardia (2002).

[Table 2 about here.]

[Figure 4 about here.]

In this case, the value of the CDC measure for the piecewise model it is 0.332 while for the c-c model is 0.394. The CDC has been reduced by 9.91%. As in the previous example the values of the GDF appearing in Table 2 are obtained by averaging the results from several runs of the GDF algorithm with different values of the tuning parameter τ. The value of the Generalized Akaike Information Criterion $GAIC$ for the piecewise model is 19.809, while for the c-c model it is 19.412, so that there is some evidence of improvement due to the piecewise model in spite of its higher complexity. The difference between these two values is not as big as might be expected from the log-likelihood values (16.448 for the piecewise model and

13.553 for the c-c model, see Table 2) as the GDF are 6.544 and 3.847 for the piecewise and the c-c model respectively. (We note that our model selection proposal is somewhat conservative as using the usual DF values, 5 for the piecewise model and 3 for c-c, gives a bigger difference (22.896 vs 21.106) between the models.) As this conservative approach still yields a difference in favor of the piecewise model we are reinforced in our conclusions. Moreover, reversing the role of the variables as we did in the previous example also yields the same result, as the $GAIC$ for the piecewise model is higher than that of the c-c model.

Another application of our method is to estimate the duration of time the cells spend in various phases of the cell-cycle. It is well-known among cell biologists that during the cell division cycle, *S. cerevisiae* spends equal time in all phases (nearly 25% in each phase) whereas *S. pombe* spends a large proportion of time (according to some estimates nearly 70%) in the G2 phase. Interestingly, our method allows us to estimate these phase durations. More precisely, the images of the change points $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*, \theta_4^*) = (2.10, 2.80, 4.00, 5.75)$ by the piecewise model are $(5.61, 6.25, 1.48, 4.76)$. Thus the lengths of the phases in the regressor species (*S. cerevisiae*) are $(0.70, 1.20, 1.75, 2.63)$, while using the piecewise model the estimated phase lengths for the response (*S. pombe*) are $(0.64, 1.51, 3.28, 0.85)$, so that there are differences in the 3rd and 4th sectors which correspond to cell phases G2 and M respectively. Thus according to our estimates, *S. cerevisiae* spends 27.85% of its time in G2 phase whereas *S. pombe* spends 52.20% of its time in the G2 phase during the cell-cycle.

4. Discussion

Piecewise regression models have proved very useful in the case of data on the line to describe multi-linear relationships representing the different effects of the explicative variable on the response, before and after some change points values on the explicative. Something similar is likely to occur in circular models. This paper is a first contribution to the study of these

models in the circular setting. The model where the change points are assumed to be known has been described, estimated and applied to cell-cycle gene expression data.

The choice of these change points is obviously an important question. In the applications we provide in Section 3, these points have been established using information available from those applications. This would certainly be the case in many applications. Otherwise, a visual inspection of the data may provide a good guide in this choice. Moreover, some other more sophisticated approaches may be used. For example, the circular isotonic regression estimator (CIRE) of a vector under a fixed order defined in Rueda et al. (2009) may be useful. Given a fixed circular order, when this estimator is computed for a vector of circular values it yields the vector that follows that circular order and is as close as possible (using a circular distance criterion, see Rueda et al. (2009)) to the original vector. Then, if we compute the CIRE of the response variable under the order given by the values of the regressor variable, we will obtain several sets of indexes (called level sets) where the CIRE is constant. A good choice for the change points would then be to select values of the regressor variable between those that generated the highest jumps among the level sets.

There are many interesting biological questions related to cell-cycle that can be answered from our circular piecewise model. For example, we can construct a formal test as to whether there is a significant difference in phase lengths in two given species. Similar to Liu et al. (2004), we can construct a test to assess if the order of the phase angles of cell-cycle genes from different labs/experiments on the same species is preserved. We can also evaluate if the phase angle has changed evolutionarily detecting these changes as the outliers in the model.

It is to be noted that the model we propose is not only interesting for cell biology applications. Another application appears when dealing with women's menstrual cycle which consists of three phases, namely, follicular, ovulation (single point) and luteal. Although the blood or urine concentration of hormones such as estrogen and progesterone are periodic

during the approximate 28-day cycle, they have distinct patterns according to the phase of the cycle. For example, during the follicular phase, the beginning of menses, the concentration of estrogen rises sharply (almost like an exponential curve) and the blood estrogen levels drop sharply to the baseline at ovulation; it then starts to rise during the luteal phase and attains its peak in the middle of the luteal phase, then begins to drop slowly towards the end of the monthly period.

The model proposed is of general interest beyond cell-cycle data. For example, it is applicable in fields such as circadian biology (Kondratova and Kondratov (2012)), evolutionary biology (De Quadros-Wander and Stokes (2007)) and motor behavior (Baayen et al. (2012)).

From a methodological point of view, there are some extensions of the proposed model that can be dealt with. One of these extensions is the inclusion of other regressor variables in the model. In the linear piecewise regression literature this is done assuming that the change points depend either on only one of the regressor variables or on the time the observations are taken (see, for example, Liu et al. (1997)). In this way, it is easy to include other circular variables $\theta_2, \ldots, \theta_s$ and/or linear ones $Z_1, \ldots, Z_t$ in the model, replacing formulation (5) by

$$\mu_{ij} = \mu + 2 \arctan \left(\sum_{l=1}^{s} \omega_{il} \tan \frac{1}{2}(\theta_{ijl} - \nu_{il}) + \sum_{m=1}^{t} \beta_{im} z_{ijm} \right),$$

where β_{im} is the slope of the linear variable Z_m in sector i, and replacing also the continuity conditions (4) by the corresponding ones. Notice that, since our estimation scheme relies on the circular-linear model from Fisher and Lee (1992), the estimation of the parameters can be performed as when only one circular regressor variable is present, although restrictions on the parameters may be more involved. The same holds for model selection. Our $GAIC$ criterion and its computation does not depend on how many variables are in the model, or if they are circular or linear, thus making easy the task of variable selection.

Another extension that can be solved easily is that of dropping the known θ_i^* assumption and estimating the phase boundaries as unknown parameters, or even the assumption of an

unknown number of sectors. As with the previous extension no change on the estimation or model selection procedure is needed. However, it is obvious that the computational burden of both parameter estimation and model evaluation will be highly increased.

5. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2.1, 2.3, 2.4, 3.1 and 3.2 are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

This work was supported by Spanish MCI grant (MTM2012-37129 to S.B., C.R. and M.A.F.), Junta de Castilla y León and the European Social Fund within the Programa Operativo Castilla y León (2007-2013 to S.B.) and the Intramural Research Program of the NIEHS (Z01 ES101744-04 to S.D.P.). The authors thank the anonymous reviewers and the associate editor for several useful comments which improved this manuscript.

References

Agostinelli, C. and Lund, U. (2011). *circular: Circular Statistics*. R package version 0.4-3.

Akaike, H. (1973). Information theory and the maximum likelihood principle. In *International Symposium on Information Theory*. Akademiai Kiádo.

Baayen, C., Klugkist, I., and Mechsner, F. (2012). A test for the analysis of order constrained hypotheses for circular data. *Journal of Motor Behavior* **44,** 351–363.

De Quadros-Wander, S. and Stokes, M. (2007). The effect of mood on opposite-sex judgments of males' commitment and females' sexual content. *Evolutionary Psychology* **4,** 453–475.

Di Marzio, M., Panzera, A., and Taylor, C. (2013). Non-parametric regression for circular responses. *Scandinavian Journal of Statistics* **40,** 238 – 255.

Downs, T. and Mardia, K. (2002). Circular regression. *Biometrika* **89,** 683–697.

Fernández, M., Rueda, C., and Peddada, S. (2012). Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Research* **40,** 2823–2832.

Fisher, N. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press.

Fisher, N. and Lee, A. (1992). Regression models for an angular response. *Biometrics* **48,** 665–677.

Kato, S. and Jones, M. (2010). A family of distributions on the circle with links to, and applications arising from, Möbius transformation. *Journal of the American Statistical Association* **105,** 249–262.

Kato, S., Shimizu, K., and Shieh, G. (2008). A circular-circular regression model. *Statistica Sinica* **18,** 633–645.

Kondratova, A. and Kondratov, R. (2012). The circadian clock and pathology of the ageing brain. *Nature Reviews Neuroscience* **13,** 325–335.

Liu, D., Weinberg, C., and Peddada, S. (2004). A geometric approach to determine association and coherence of the activation times of cell-cycling genes under differing experimental conditions. *Bioinformatics* **20,** 2521–2528.

Liu, J., Wu, S., and Zidek, J. (1997). On segmented multivariate regression. *Statistica Sinica* **7,** 497–525.

Mardia, K. and Jupp, P. (2000). *Directional Statistics*. John Wiley & Sons.

Muggeo, V. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics* **27,** 161–166.

Muller, S., Scealy, J., and Welsh, A. (2013). Model selection in linear mixed models. *Statistical Science* **28,** 135–167.

Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., et al. (2005). The cell-cycle-regulated genes of schizosaccharomyces pombe. *PLoS Biology* **3,** 1239–1260.

Painting, C. and Holwell, G. (2013). Exaggerated trait allometry, compensation and trade-offs in the New Zealand giraffe weevil (Lasiorhynchus barbicornis). *PLoS ONE* **8,** e82467.

Polsen, O. and Taylor, C. (2015). Parametric circular-circular regression and diagnostic analysis. In Dryden, I. and Kent, J., editors, *Geometry Driven Statistics*. Wiley.

Pramila, T., Wu, W., Miles, S., Noble, W., and Breeden, L. (2006). The forkhead transcription factor hcm1 regulates chromosome segregation genes and fills the s-phase gap in the transcriptional circuitry of the cell cycle. *Genes and Development* **22,** 2266–2278.

Rueda, C. (2013). Degrees of freedom and model selection in semiparametric additive monotone regression. *Journal of Multivariate Analysis* **117,** 88–99.

Rueda, C., Fernández, M., and Peddada, S. (2009). Estimation of parameters subject to order restrictions on a circle with application to estimation of phase angles of cell-cycle genes. *Journal of the American Statistical Association* **104,** 338–347.

Santos, A., Wernersson, R., and Jensen, L. (2015). Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research* **43,** D1140–D1144.

Seber, G. and Wild, C. (1989). *Nonlinear Regression*. John Wiley and Sons.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9,** 3273–3297.

Tibshirani, R. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **40,** 1198–1232.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93,** 120 – 131.

Zhang, B., Shen, X., and Mumford, S. (2012). Generalized degrees of freedom and adaptive model selection in linear mixed-effects models. *Computational Statistics and Data Analysis* **56,** 574 – 586.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

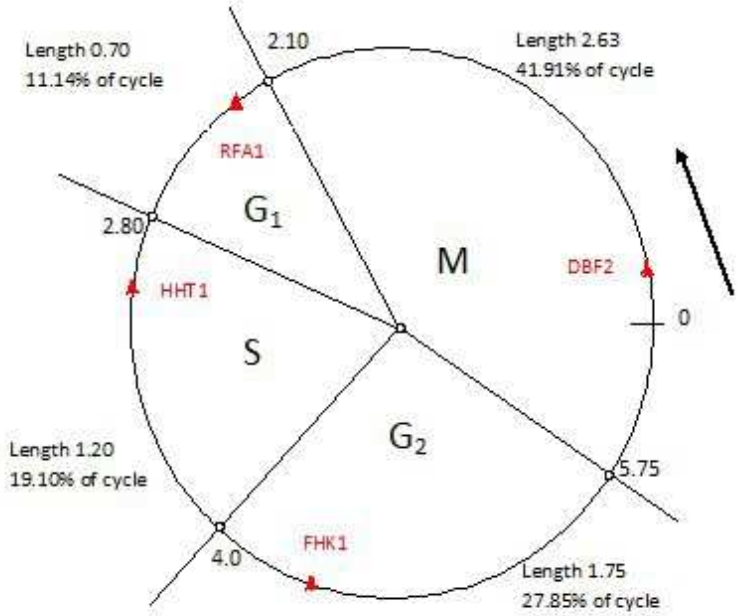

Figure 1. The four phases of a normal cell division cycle (G1, S, G2 and M) together with their sector boundaries, phase length and their relative percentages of the time spent. The arrow shows the direction of the cell-cycle. Four data points (the phase angle of the four genes RFA1, HHT1, FHK1, and DBF2) are also displayed (by triangles).

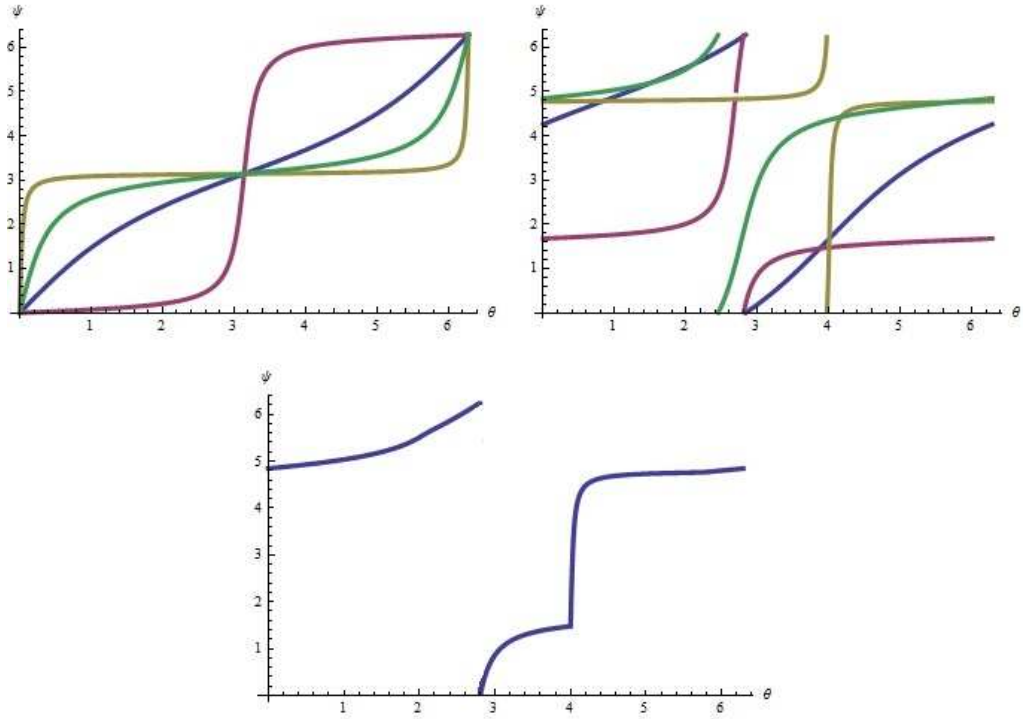

Figure 2. Top left: curves used in the piecewise model; Top right: curves used in the piecewise model shifted to their actual location; Bottom: final regression curve.

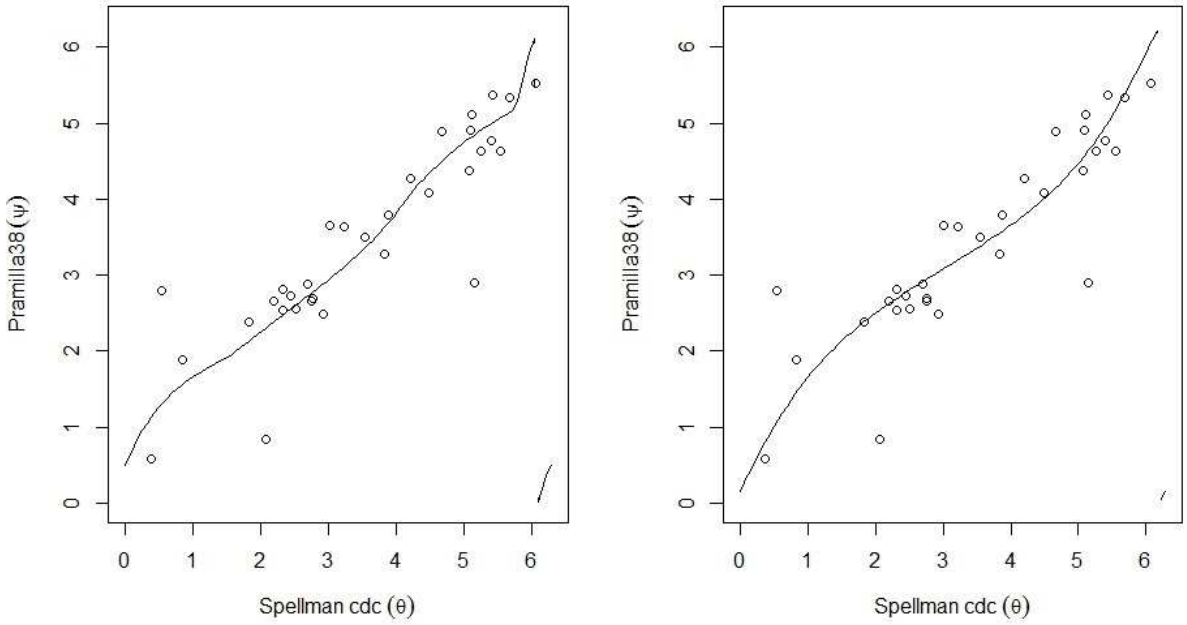

Figure 3. Fit of the estimated models for the *S. Cerevisiae* data (piecewise at left and c-c at right).

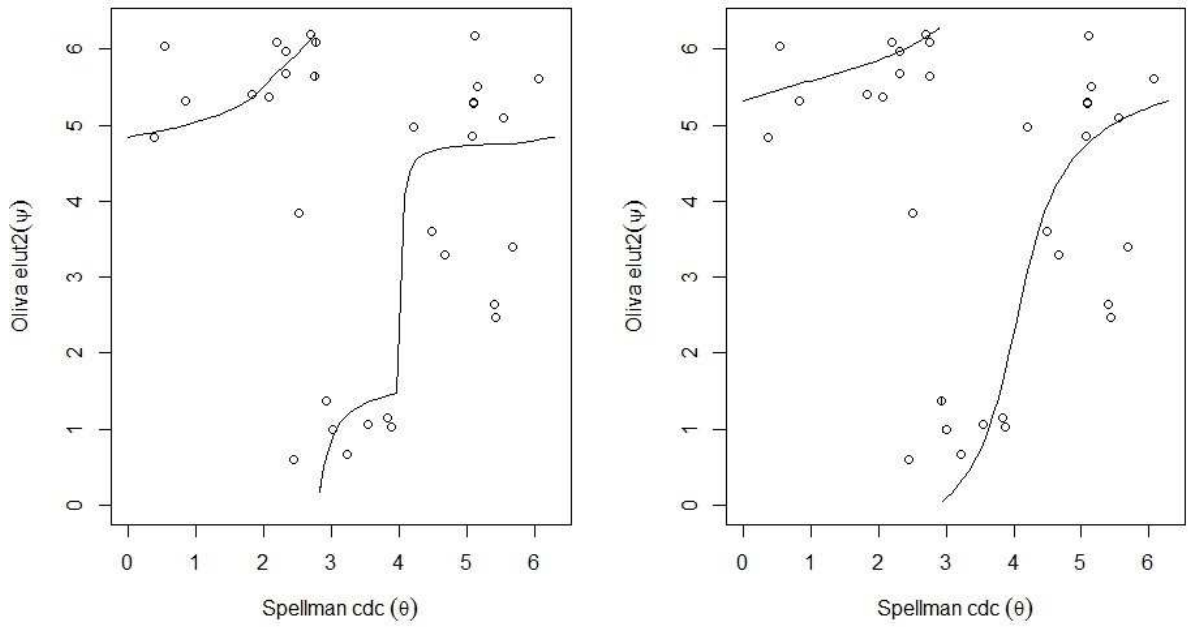

Figure 4. Fit of the estimated models for the two species data (piecewise at left and c-c at right)

Table 1
Maximum likelihood estimates and diagnostics under the piecewise regression model and the c-c model for the S. Cerevisiae data

Parameters	Estimated values (piecewise)	Estimated values (c-c)
μ	-0.840	3.099
ω	(1.491, 1.664, 0.485, 3.416)	0.546
ν	(5.214, 5.138, 0.003, 5.828)	3.032
κ	3.734	3.511
CDC	0.138	0.148
$\ln(l(m))$	32.723	31.421
GDF	4.986	3.432
$GAIC$	55.474	55.977

Table 2
Parameter estimations and diagnostics obtained using the piecewise regression model and the c-c model for the two species data

Parameters	Estimated values (piecewise)	Estimated values (c-c)
μ	1.646	-0.725
ω	(1.658, 0.066, 65.711, 6.517)	0.244
ν	(3.986, 5.824, 4.003, 2.774)	0.897
κ	1.822	1.592
CDC	0.332	0.394
$\ln(l(M))$	16.448	13.553
GDF	6.544	3.847
$GAIC$	19.809	19.412