**University of São Paulo
“Luiz de Queiroz” College of Agriculture**

**Estimating the longitudinal concordance correlation through fixed
effects and variance components of polynomial mixed-effects
regression model**

Thiago de Paula Oliveira

Thesis presented to obtain the degree of Doctor in Science
Area: Statistics and Agricultural Experimentation

**Piracicaba
2018**

Thiago de Paula Oliveira
Agronomy Engineer

**Estimating the longitudinal concordance correlation through fixed
effects and variance components of polynomial mixed-effects
regression model**

Orientador:

Prof. Dr. **SILVIO SANDOVAL ZOCCHI**

Thesis presented to obtain the degree of Doctor in Science
Area: Statistics and Agricultural Experimentation

Piracicaba
2018

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP**

Oliveira, Thiago de Paula

Estimating the longitudinal concordance correlation through fixed effects and variance components of polynomial mixed-effects regression model / Thiago de Paula Oliveira. – Piracicaba, 2018 .

115 p.

Tese (Doutorado) – USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Dados longitudinais 2. Modelo de regressão linear misto 3. Concordância longitudinal 4. *Carica papaya* L. 5. Análise de cor 6. Software R .
I. Título.

DEDICATORY

To my parents José Aparecido and Rosângela,
And my fiancé Letícia.
I love you.

ACKNOWLEDGMENTS

I thank to the University of São Paulo, especially “Luiz de Queiroz” College of Agriculture (ESALQ) for my professional qualification and National University of Ireland, that supported this research project. Financial support from the CNPq (“Conselho Nacional de Desenvolvimento Científico e Tecnológico”); and the CAPES (“Coordenação de Aperfeiçoamento de Pessoal de Nível Superior”) is gratefully acknowledged. Additional support was provided by FAPESP (“Fundação de Amparo à Pesquisa do Estado de São Paulo”) and by Departamento de Ciências Exatas of ESALQ.

Special acknowledgment goes to my advisor Prof. Silvio Sandoval Zocchi, who has always believed in my potential, and to my supervisor abroad John Hinde for the support of my Ph.D thesis and for they academic and friendly advices. I am also very thankful to my friends from Departamento de Ciências Exatas (ESALQ/USP) and from Statistical Genetics Laboratory (ESALQ/USP).

I would like to extend my acknowledgments to my family (Cido, Rose, and Tati) and my family in law (José Maurício, Liberaci, Renato, Bruna, Lais, and Gabriel), for their unconditional support during all stages of my life and for always believing in my potential. I could not forget to mention my other two families Vic, Ronaldo, Renan, Marina and Enzo, and Estela, Ronaldo, Rafael, Marcus, Ana, and Bruno that will always be kept in my heart.

I also need to thank my friends and siblings: Silvio Zocchi for his support, friendship, and confidence; Idemauro Lara, who became a great friend and part of my family; Rafael Moral, with whom I have always shared every stage of my life; and Aline, Adriano, Isabela and Fábio, who are loyal friends.

Finally, I would like to thank my fiancé Letícia Aparecida de Castro Lara for all comprehension, attention and love that she had with me through all these years.

“Live as if you were to die tomorrow. Learn as if you were to live forever.”

Mahatma Gandhi

Até a calmaria chegar

Eu estou sozinho
Não posso nem mesmo imaginar
As incertezas do caminho
De um barco perdido em alto mar

Onde está a estrela que há tanto tempo procuro?
Às vezes fico assim pensando...
Em episódios arrítmicos
E, assim, enlouquecendo a procura de um lar!

Estou completamente perdido nessa imensidão
Como posso, então, fixar meu olhar
Em estrelas que me confundem o tempo todo?
São apenas portas abertas aos tolos sob a luz do luar!

Lá se vai meu pequeno barco novamente
Levado por outra onda doutro amor luzente
Até a calmaria chegar!

Thiago de Paula Oliveira

SUMMARY

Resumo	8
Abstract	9
1 Introduction	11
2 Conclusions	12
References	13

RESUMO

Estimando a correlação de concordância longitudinal por meio de efeitos fixos e componentes de variâncias do modelo de regressão polinomial de efeitos mistos

Resumo: No setor de pós-colheita é muito comum a utilização de colorímetros para avaliar a cor média da casca de frutos ao longo do tempo. No entanto, muitas vezes as técnicas de amostragem utilizando esse equipamento pode levar a medidas tendenciosas da média amostral. Alternativamente, a utilização de imagens digitais podem levar a um menor viés, uma vez que toda a região da casca do fruto é amostrada de forma sistemática. No entanto, ainda é necessária a comparação de ambas abordagens, pois o colorímetro tem vantagens em relação a facilidade de utilização e menor tempo para realizar a amostragem em cada fruto quando comparado a um scanner de mesa. Assim, no caso de variáveis respostas medidas em uma escala contínua, a reprodutibilidade das medidas tomadas por ambos equipamentos pode ser avaliada por meio do coeficiente de correlação de concordância. Dessa forma, para avaliar o perfil da concordância entre métodos nós propomos uma correlação de concordância longitudinal (LCC), baseada em um modelo de regressão polinomial com efeitos mistos. Os resultados sugeriram que as técnicas por meio de imagens digitais devem ser utilizadas para a quantificação da tonalidade média de frutos. Adicionalmente, a partir do perfil de concordância estimado notamos que existe um período em que ambos os equipamentos podem ser utilizados. A performance do coeficiente de concordância longitudinal foi avaliada por meio de um estudo de simulação, o qual sugeriu que nossa metodologia é robusta a dados desbalanceados (“dropout”) e que a probabilidade de convergência é aceitável para uma amostra de 20 frutos e ideal para amostras a partir de 100 frutos. Uma vez que ainda não existem pacotes disponibilizados no ambiente computacional R para a estimação da correlação de concordância longitudinal, nós estamos desenvolvendo um pacote intitulado `lcc`, o qual será submetido ao “comprehensive R archive network” (CRAN). Nesse pacote nós implementamos procedimentos para estimação da correlação de concordância longitudinal, da correlação de Person longitudinal e de uma medida de acurácia longitudinal. Além disso, nosso pacote suporta dados balanceados e desbalanceados, permite modelar a heteroscedasticidade entre erros dentro do grupo usando ou não o tempo como covariável, e também permite a inclusão de covariáveis no preditor linear para controlar variações sistemáticas na variável resposta.

Palavras-chave: Dados longitudinais, Modelo de regressão linear misto, Concordância longitudinal, *Carica papaya* L., Análise de cor, Software R

ABSTRACT

Estimating the longitudinal concordance correlation through fixed effects and variance components of polynomial mixed-effects regression model

In the post-harvest area, a common approach to quantify the average color of fruits peel over time is the sampling of small number of points generally on its equatorial region using a colorimeter. However, when we use a colorimeter to classify an uneven-colored fruit misclassification may occur because points in the peel region may not be representative of average color of fruit. The main problem when we use this method is to determine the number of points to be sampled as well as the location of these points on the fruit's surface. An alternative method to evaluate measure of color is digital image analysis because it covers whole of the object surface, by using a sample of pixels taken from the image. As the colorimeter approach is faster and easier than image analysis, it may not be suitable for assessing the overall mean color of the papaya's peel and its performance will depend on the number of measured points and choice of sampled region. In this sense, the comparison between these approach is still necessary because we need to know if a sample on the equatorial region can reproduce a sample over the whole region, and if the colorimeter can compete with a scanner or digital camera in measuring the mean hue of papaya peel over time. Thus, we proposed a longitudinal concordance correlation (LCC) based on polynomial mixed-effects regression model to evaluate the extent of agreement among methods. The results show that ideally image analysis of whole fruit's region should be used to compute the mean hue and that the topography and curved surface of papaya fruit did not affect the mean hue obtained by the scanner. Since there are still no packages available to estimate the LCC in the free software environment R, we are developing a package called *lcc*, which provides functions for estimating the longitudinal concordance correlation (LCC) among methods based on variance components and fixed effects of polynomial mixed-effects model. Additionally, we implemented arguments in this function to estimating the longitudinal Pearson correlation (LPC), as precision measure, and longitudinal bias corrector factor (LA), as accuracy measure. Moreover, these components can be estimated using different structures for variance-covariance matrices of random effects and variance functions to model heteroscedasticity among within-group errors using or not the time as variance covariate.

Keywords: Longitudinal data, Mixed-effects regression, Longitudinal agreement, *Carica papaya* L., Color analysis, Software R

1 INTRODUCTION

Measuring the amount of agreement between two or more responses in a common research goal. When the data are categorical in nature, the Cohen’s kappa (COHEN, 1960) and weighted kappa (COHEN, 1968) statistics are appropriated measures of agreement for ordinal and nominal categorical variables, respectively. On the other hand, when the data are measured in continuous scale, the concordance correlation coefficient (CCC) introduced by LIN (1989) should be used as agreement measure between two methods. This agreement index is calculated by measuring the variation of the linear relationship between each pair of observations from the 45° line through the origin. Therefore, the mainly advantage of this coefficient is not only measuring how far each observation deviates from the best-fit line (measure of precision), but also how far the best-fit line deviates from the 45° line through the origin (measure of accuracy). Moreover, some extensions of CCC to categorical data was developed to estimates agreement equivalently to Cohen’s kappa and weighted kappa statistics (KRIPPENDORFF, 1970; KING and CHINCHILLI, 2001).

However, when subjects, methods, or covariates variation are considered in the estimation of CCC, CARRASCO and JOVER (2003) suggested the use of two-way linear mixed-effect model to estimate CCC through variance components (VC), which was extended to repeated measures by CARRASCO *et al.* (2009) and soon after to generalized linear-effects models (GLMMs) for count data by CARRASCO (2010). Other research direction is to estimated CCC via generalized estimating equations (GEE) introduced by BARNHART and WILLIAMSON (2001), who proposed three sets of estimating equations to model CCC between two methods for continuous variables. This approach was also extended to multiple methods (BARNHART *et al.*, 2005). Additionally, TSAI (2015) demonstrated through simulation study that CCC under VC and GEE can be biased when the model is misspecified, consequently, it is very important to perform model selection in both approach. Besides, this author also show that VC approach is less affected by model misspecification than GEE one.

In longitudinal studies where multiple methods measure the same individual and the main goal is to evaluate whether the methods agree with each other, it is natural to want to know when interchangeably among them will be possible. This would allows to use of cheapest or most convenient method in place of others. Thus, in this thesis we proposed an extension of ideas presented by LIN (1989) and CARRASCO *et al.* (2009) based on the mixed-effects polynomial regression model rather than a mixed-effects ANOVA approach. Our approach allows to estimate the extent of agreement among pair of observations for multiple methods. Furthermore, we also proposed a R (R CORE TEAM, 2017) package called `lcc`, which provides functions for estimating the longitudinal concordance correlation (LCC) among methods based on variance components and fixed effects of polynomial mixed-effects model. We also adding arguments in this function to estimating the longitudinal Pearson correlation (LPC), as precision measure, and longitudinal bias corrector factor (LA), as accuracy measure.

2 CONCLUSIONS

In this thesis, we aimed to develop and extend statistical methods to help in the concordance analysis of agricultural data. Initially, we developed a longitudinal concordance correlation (LCC), longitudinal Pearson correlation (LPC), and longitudinal accuracy (LA) based on fixed effects and variance components of a polynomial mixed-effects regression model. We also developed methods to compute bootstrap-based confidence intervals for all components. In a small simulation study we demonstrated a satisfactory performance for LCC. The use of our approach in a real data application suggested that image analysis of whole peel region of papaya's fruit should be used to compute the mean hue rather than colorimeter or images from equatorial region.

We have been developing an R package called `lcc` for estimating LCC, LPC, and LA with or without covariates. The package allows for inclusion of heteroscedasticity among within-group error, different structures for variance-covariance matrix of random effects $\mathbf{G}$, and bootstrap-based procedure to build confidence intervals. Moreover, numerical and graphical summaries of these statistics were included in order to facilitate visualization and interpretations of results.

Further work includes extensions to the LCC approach in order to accounts for the longitudinal and spacial correlation among residuals. Improvements on the package developed here are still subject of ongoing research, and we appreciate any suggestion. Another extension would be a simulation study to compare several type of bootstrap confidence intervals (parametric and non-parametric). Finally, we would be happy to share the code of our package, which is not already available on CRAN.

References

- BARNHART, H., J. SONG, and M. HABER, 2005 Assessing intra, inter, and total agreement with replicated measurements. *Stat. Med.* **24**: 1371–1384.
- BARNHART, H. X. and J. M. WILLIAMSON, 2001 Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* **57**: 931–940.
- CARRASCO, J. L., 2010 A generalized concordance correlation coefficient based on the variance components generalized linear mixed models for overdispersed count data. *Biometrics* **66**: 897–904.
- CARRASCO, J. L. and L. JOVER, 2003 Estimating the Generalized Concordance Correlation Coefficient through Variance Components. *Biometrics* **59**: 849–858.
- CARRASCO, J. L., T. S. KING, and V. M. CHINCHILLI, 2009 The concordance correlation coefficient for repeated measures estimated by variance components. *Journal of biopharmaceutical statistics* **19**: 90–105.
- COHEN, J., 1960 A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**: 37–46.
- COHEN, J., 1968 Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**: 213–220.
- KING, T. S. and V. M. CHINCHILLI, 2001 A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* **20**: 2131–2147.
- KRIPPENDORFF, K., 1970 Bivariate Agreement Coefficients for Reliability of Data. *Sociological Methodology* **2**: 139–150.
- LIN, L. I., 1989 A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **45**: 255–268.
- R CORE TEAM, 2017 The R environment.
- TSAI, M.-Y., 2015 Comparison of concordance correlation coefficient via variance components, generalized estimating equations and weighted approaches with model selection. *Computational Statistics and Data Analysis* **82**: 47–58.