Imputação de dados climáticos e de produtividade agrícola: Uma comparação de abordagens

Ramiro Ruiz Cárdenas (ramiro@est.ufmg.br)

Elias Teixeira Krainski (eliaskr@ufpr.br)

Marcelo Azevedo Costa (azevedo@est.ufmg.br)

1º Workshop do projeto PROCAD:

Seguro Agrícola: Modelagem Estatística e Precificação

UFMG - Belo Horizonte, Novembro 25-27 de 2009

Outline of the talk

- Motivation
- Sources of data
- Imputation of crop yield series
- Imputation of weather data
    - temperature
    - precipitation
- Data interpolation
- Future work

Outline of the talk

- Motivation
- Sources of data
- Imputation of crop yield series
- Imputation of weather data
  - temperature
  - precipitation
- Data interpolation
- Future work

Outline of the talk

- Motivation
- Sources of data
- Imputation of crop yield series
- Imputation of weather data
  - temperature
  - precipitation
- Data interpolation
- Future work

Outline of the talk

- Motivation
- Sources of data
- Imputation of crop yield series
- Imputation of weather data
  - temperature
  - precipitation
- Data interpolation
- Future work

Outline of the talk

- Motivation
- Sources of data
- Imputation of crop yield series
- Imputation of weather data
  - temperature
  - precipitation
- Data interpolation
- Future work

Outline of the talk

- Motivation
- Sources of data
- Imputation of crop yield series
- Imputation of weather data
  - temperature
  - precipitation
- Data interpolation
- Future work

Outline of the talk

- Motivation
- Sources of data
- Imputation of crop yield series
- Imputation of weather data
  - temperature
  - precipitation
- Data interpolation
- Future work

Outline of the talk

- Motivation
- Sources of data
- Imputation of crop yield series
- Imputation of weather data
  - temperature
  - precipitation
- Data interpolation
- Future work

Motivation

- The parameters for a weather based insurance contract are generally derived from historical weather data. Without an appropriate quantity of relevant, high quality data, pricing and management of weather risk would be unfeasible.
- Weather data are usually subject to different types of errors (missing observations, unreasonable readings, spurious zeroes, etc.), which must be cleaned in order to be used in pricing and risk management.
- Decision support systems based on crop simulation models also rely heavily on "clean" weather data.
- Drought monitoring programs and extreme event hydrological studies also depend on reliable long term weather data series.

Motivation

- The parameters for a weather based insurance contract are generally derived from historical weather data. Without an appropriate quantity of relevant, high quality data, pricing and management of weather risk would be unfeasible.
- Weather data are usually subject to different types of errors (missing observations, unreasonable readings, spurious zeroes, etc.), which must be cleaned in order to be used in pricing and risk management.
- Decision support systems based on crop simulation models also rely heavily on "clean" weather data.
- Drought monitoring programs and extreme event hydrological studies also depend on reliable long term weather data series.

Motivation

- The parameters for a weather based insurance contract are generally derived from historical weather data. Without an appropriate quantity of relevant, high quality data, pricing and management of weather risk would be unfeasible.
- Weather data are usually subject to different types of errors (missing observations, unreasonable readings, spurious zeroes, etc.), which must be cleaned in order to be used in pricing and risk management.
- Decision support systems based on crop simulation models also rely heavily on “clean” weather data.
- Drought monitoring programs and extreme event hydrological studies also depend on reliable long term weather data series.

Motivation

- The parameters for a weather based insurance contract are generally derived from historical weather data. Without an appropriate quantity of relevant, high quality data, pricing and management of weather risk would be unfeasible.
- Weather data are usually subject to different types of errors (missing observations, unreasonable readings, spurious zeroes, etc.), which must be cleaned in order to be used in pricing and risk management.
- Decision support systems based on crop simulation models also rely heavily on “clean” weather data.
- Drought monitoring programs and extreme event hydrological studies also depend on reliable long term weather data series.

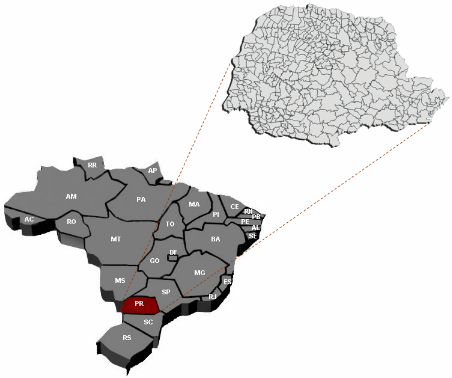
Study region and available data sets

- Crop yield data:
  average annual county yield (1980 – 2007).
  source: IBGE / SEAB
  http://www.sidra.ibge.gov.br

- Meteorological data:

  daily precipitation series for 503 stations (01/01/76 – 31/12/08).
  source: ANA / SUDHERSA / IAPAR / SIMEPAR / INMET
  http://hidroweb.ana.gov.br

  daily temperature series for 87 stations (01/01/76 – 31/12/08).
  source: INMET / IAPAR / SIMEPAR

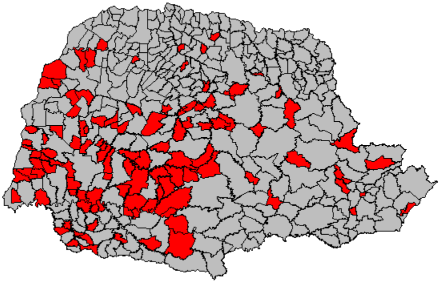


State: Paraná
N^{o} counties: 399
planted area (grains): 8.45 mill Ha

Recovering the crop yield time series

- 109 counties were created between 1983 and 1997 from existing ones.

year	counties
1983	20
1986	1
1989	7
1990	5
1993	48
1997	28



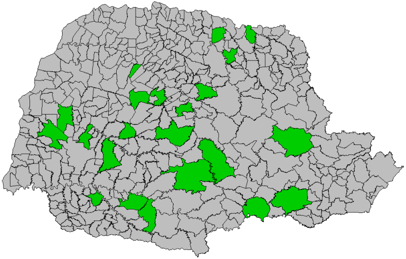
* source: IBGE 2009

Recovering the crop yield time series

A simulation study:

- some counties and its neighbors with complete yield series (1980-2008) were used to simulate the creation of new counties

- Nº of created counties: 22
- years of creation:
  1983
  1987
  1992
  1997
- Former counties:
  - best correlated neighbors
  - worst correlated neighbors



Recovering the crop yield time series

$$joint.area = area[old, after] + area[new, after]$$
$$joint.pdn = pdn[old, after] + pdn[new, after]$$

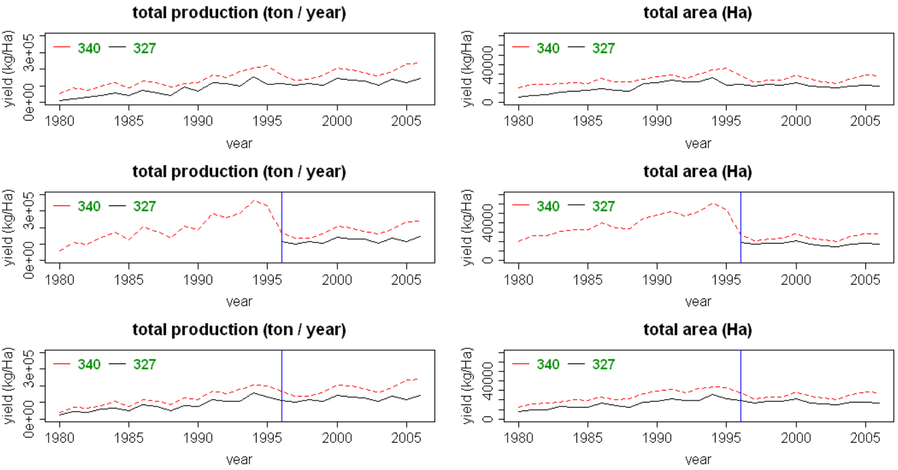
$$prop.area.new = \frac{area[new, after]}{joint.area} \quad ; \quad prop.pdn.new = \frac{pdn[new, after]}{joint.pdn}$$

$$(a, b) = mean(prop.area.new[1 : {\color{red}w}]) \pm {\color{red}k} * sd(prop.area.new[1 : {\color{red}w}])$$
$$(c, d) = mean(prop.pdn.new[1 : {\color{red}w}]) \pm {\color{red}k} * sd(prop.pdn.new[1 : {\color{red}w}])$$

$$prop.area.new.before = runif({\color{red}nn}, a, b)$$
$$prop.pdn.new.before = runif({\color{red}nn}, c, d)$$

$$yield[new, before] = \frac{pdn[old, before] * prop.pdn.new.before}{area[old, before] * prop.area.new.before}$$

Recovering the crop yield time series



Recovering the crop yield time series

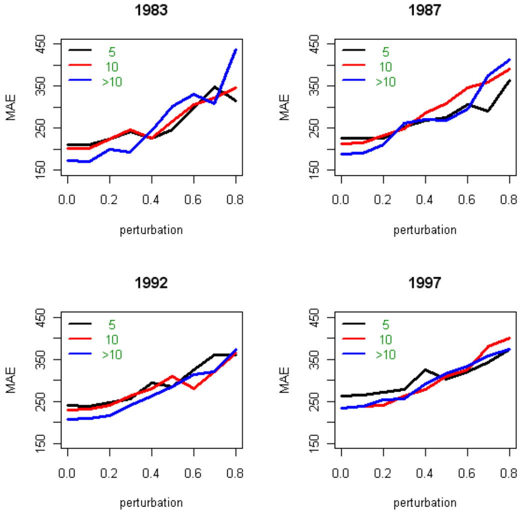


Figure 1. Mean absolute error for all the scenarios applied on corn yield series simulated from the best correlated neighbors.

Recovering the crop yield time series

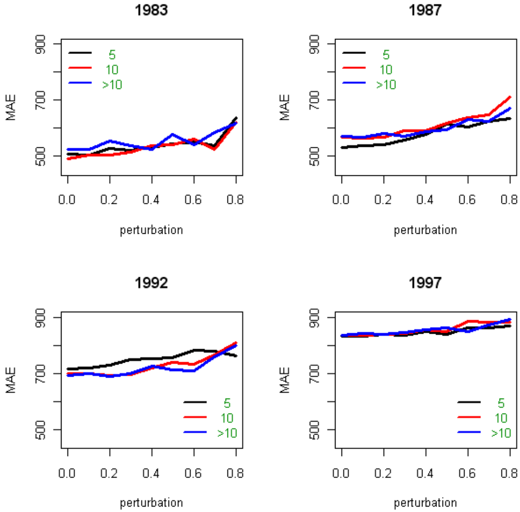


Figure 2. Mean absolute error for all the scenarios applied on corn yield series simulated from the worst correlated neighbors.

Recovering the crop yield time series

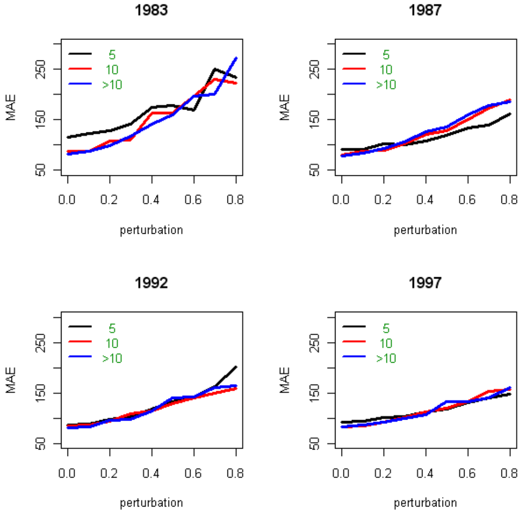


Figure 3. Mean absolute error for all the scenarios applied on soybean yield series simulated from the best correlated neighbors.

Recovering the crop yield time series

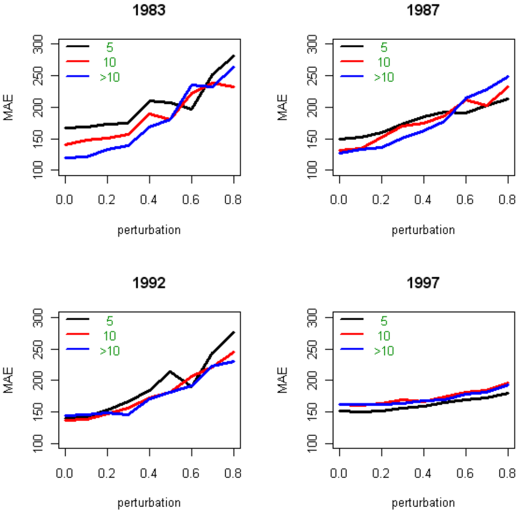


Figure 4. Mean absolute error for all the scenarios applied on soybean yield series simulated from the worst correlated neighbors.

Recovering the crop yield time series

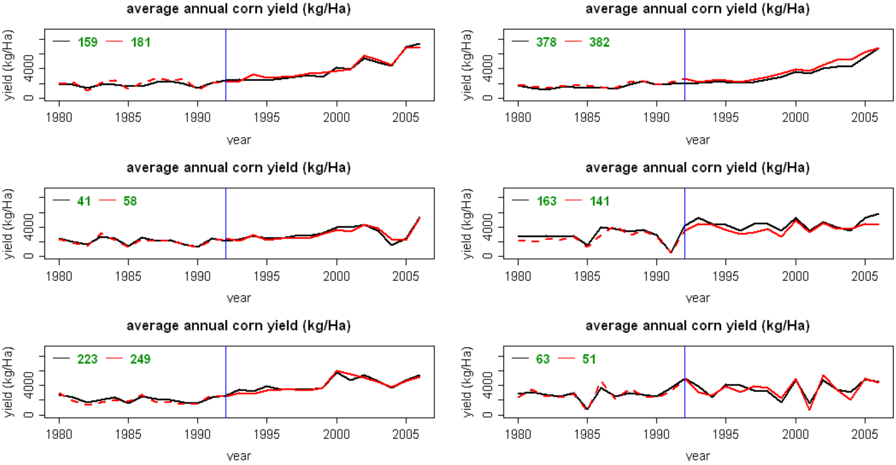


Figure 5. Corn yield series recovered in six new counties from its parents

Imputing the weather data

Weather variables to be imputed:

- minimum temperature
- maximum temperature
- precipitation

Temporal scales:

- daily (12054 values/station)
- decendial (1188 values/station)

Imputation approaches:

- EM algorithm (Junger et al., 2003, Schneider, 2001)
- Principal component analysis (Stacklies, 2007)
- Multiple imputation (Van Buuren, 2006)
- Neural Networks (Kim et al., 2009)
- Regression based approaches

Imputing the weather data

Weather variables to be imputed:

- minimum temperature
- maximum temperature
- precipitation

Temporal scales:

- daily (12054 values/station)
- decendial (1188 values/station)

Imputation approaches:

- EM algorithm (Junger et al., 2003, Schneider, 2001)
- Principal component analysis (Stacklies, 2007)
- Multiple imputation (Van Buuren, 2006)
- Neural Networks (Kim et al., 2009)
- Regression based approaches

Imputing the weather data

Weather variables to be imputed:

- minimum temperature
- maximum temperature
- precipitation

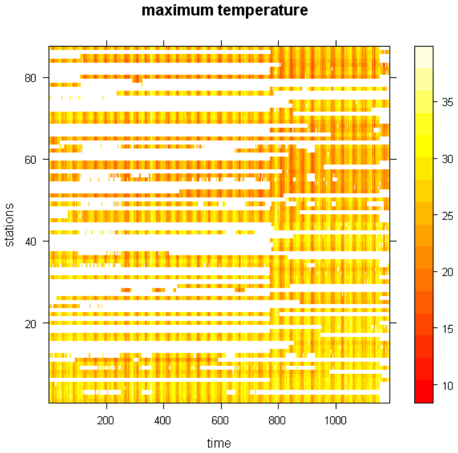
Temporal scales:

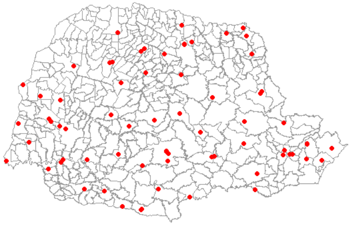
- daily (12054 values/station)
- decendial (1188 values/station)

Imputation approaches:

- EM algorithm (Junger et al., 2003, Schneider, 2001)
- Principal component analysis (Stacklies, 2007)
- Multiple imputation (Van Buuren, 2006)
- Neural Networks (Kim et al., 2009)
- Regression based approaches

Imputing the weather data





- 87 weather stations
- 33 years of records (1976 – 2008)
- 45% of missing values
- 56% of the series have < 30% of data

Methodology

- 30 different scenarios were created from the combination of the following factors:
  - 3 weather variables (Tmin, Tmax, rainfall)
  - 2 temporal scales (daily, decendial)
  - 5 sizes of subsets of observed values to be removed (1 month, 3 months, 6 months, 1 year, 3 years)
- 20 subsets of observed values were removed from each scenario and then imputed according to six imputation methods
- 5 criteria were used to compare the performance of the imputation methods.

Example:
Scenario: 1
variable: minimum temperature
temporal scale: daily
subsets to be removed/imputed: 20 subsets of 1 month each

Comparison criteria

- RMSE: Root mean square error
- MAE: Mean absolute error
- MRE: Mean relative error
- SRD: Standard deviation of the relative differences between known and imputed values

$$RD_{ij} = \frac{|Y_{ij}.obs - Y_{ij}.imp|}{|Y_{ij}.obs|} \qquad MRD = \frac{1}{m}\sum_{i \in M} RD_{ij} \qquad SRD = \sqrt{\frac{1}{m}\sum_{i \in M}(RD_{ij} - MRD)^2}$$

- MRZ: Mean number of SRD's by which a relative difference deviates from the its mean value

$$RZ_{ij} = \frac{RD_{ij} - MRD}{SRD} \qquad MRZ = \frac{1}{z}\sum_{i \in Z} RZ_{ij}$$

Imputation approaches

Multiple imputation

- MICE

- Amelia

Imputation approaches

Principal component analysis

- Probabilistic PCA
- Bayesian PCA

Imputation approaches

EM algorithm

- mtsdi

- Regularized EM

Preliminary results

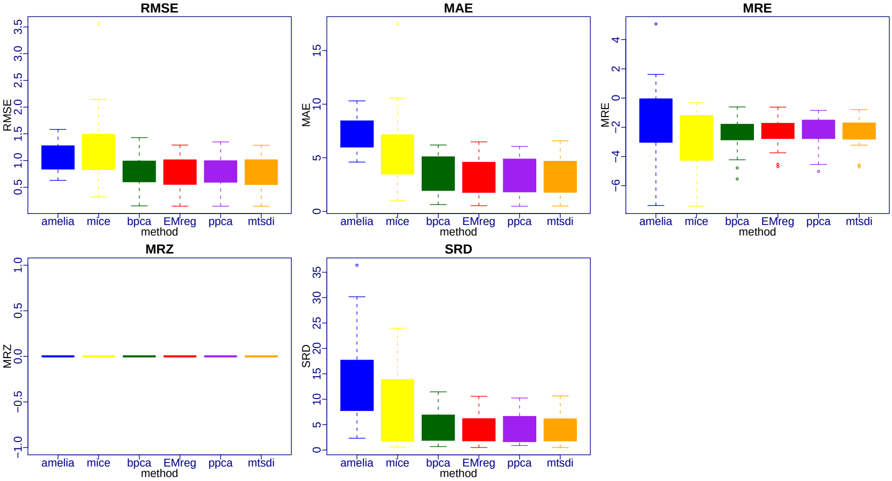


Figure 6. Boxplots for scenario 1 (daily rainfall and removing 20 periods of 3 months).

Preliminary results

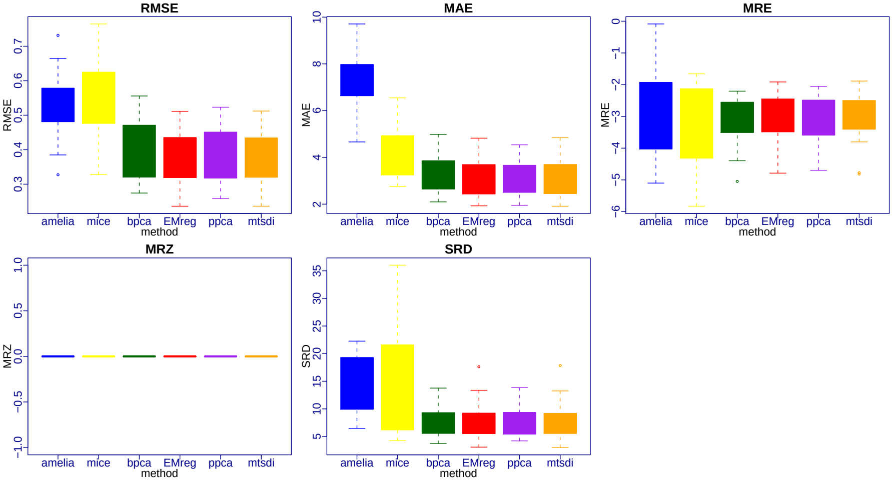


Figure 7. Boxplots for scenario 2 - (daily rainfall and removing 20 periods of 12 months).

Preliminary results

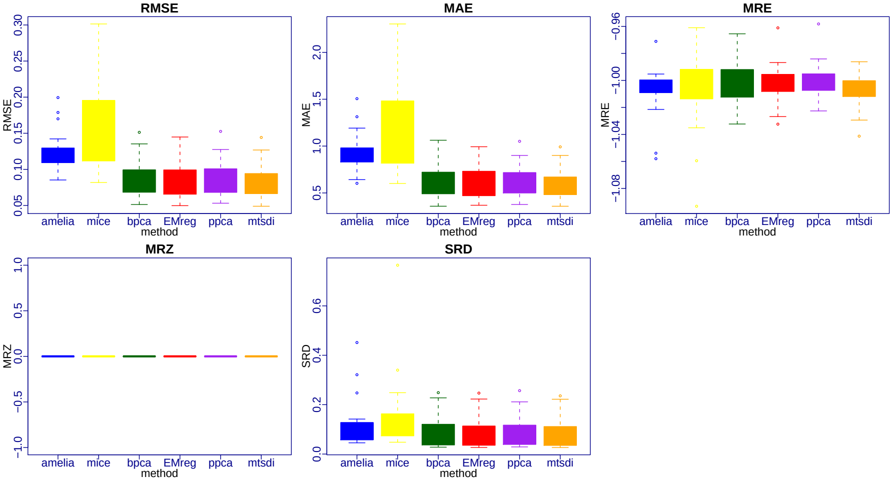


Figure 8. Boxplots for scenario 3 - (daily minimum temperature and removing 20 periods of 3 months).

Preliminary results

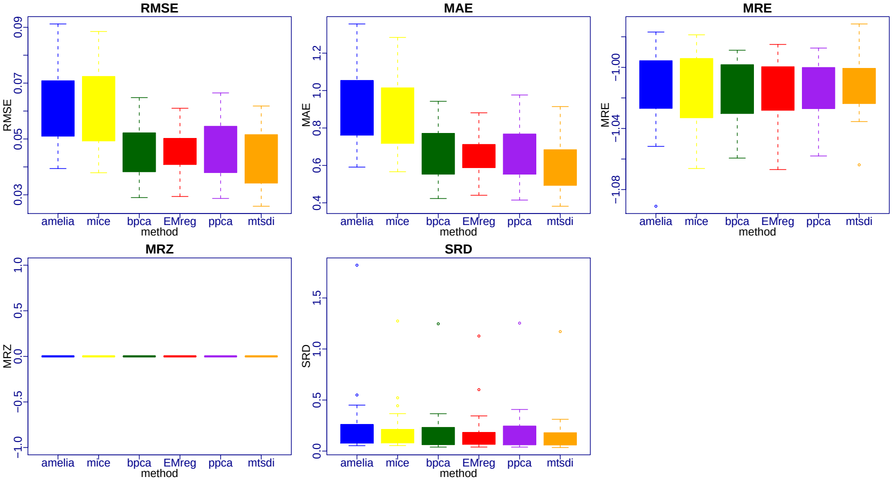


Figure 9. Boxplots for scenario 4 - (daily minimum temperature and removing 20 periods of 12 months).

Preliminary results

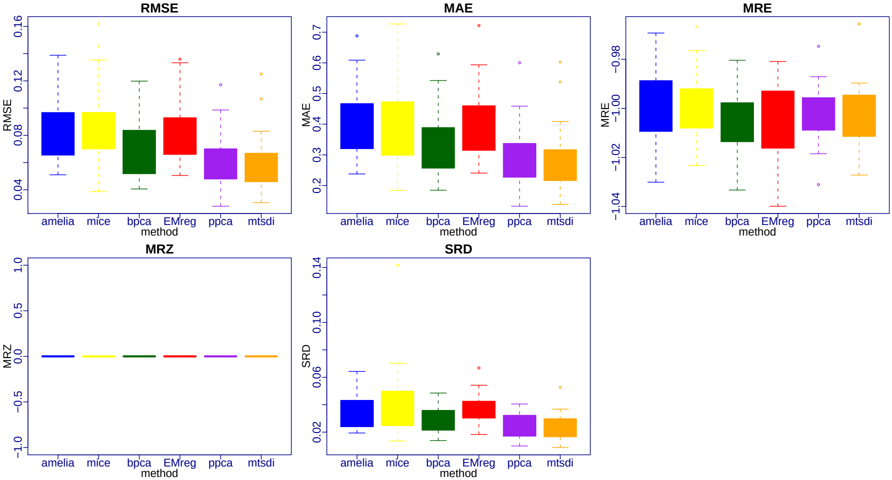


Figure 10. Boxplots for scenario 5 - (decendial minimum temperature and removing 20 periods of 12 months).

Preliminary results

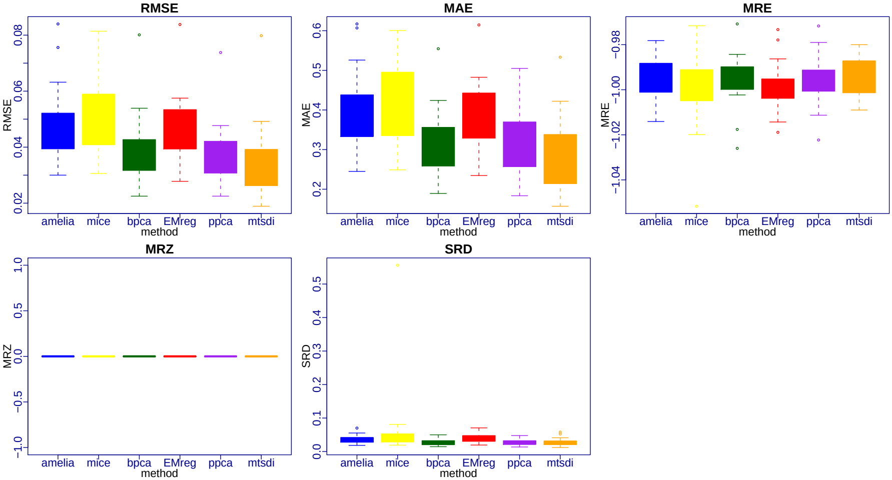


Figure 11. Boxplots for scenario 6 - (decendial minimum temperature and removing 20 periods of 36 months).

Preliminary results

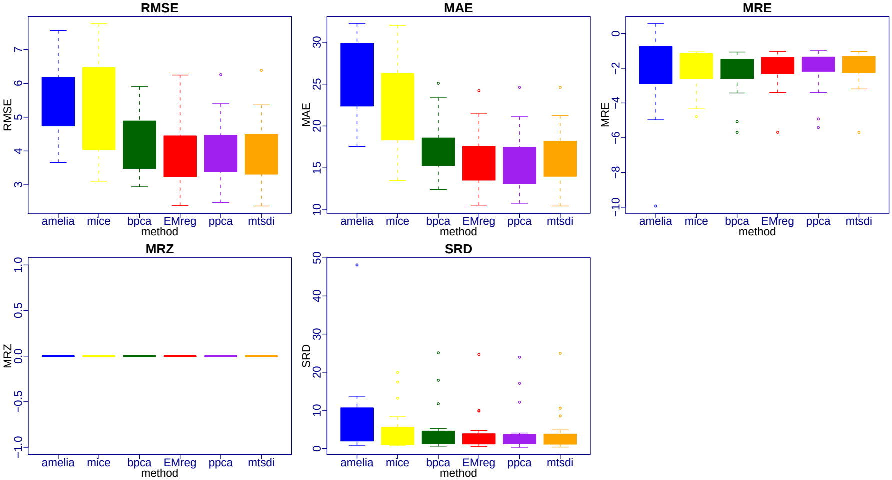


Figure 13. Boxplots for scenario 7 - (decendial rainfall and removing 20 periods of 12 months).

Preliminary results

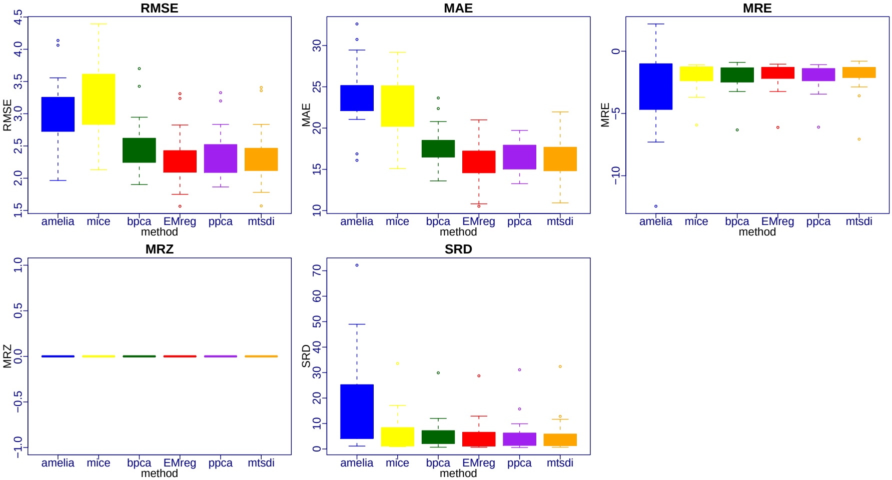


Figure 14. Boxplots for scenario 8 - (decendial rainfall and removing 20 periods of 36 months).

Preliminary results

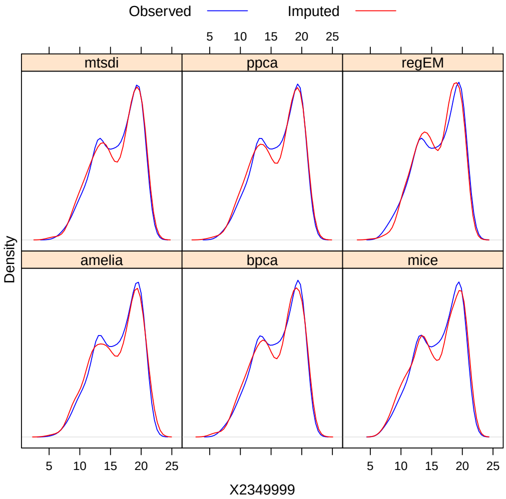


Figure 15. Kernel density estimates for the marginal distributions of the observed and imputed values at station X2349999 under scenario xx - (decendial minimum temperature and removing 20 periods of 12 months).

Preliminary results

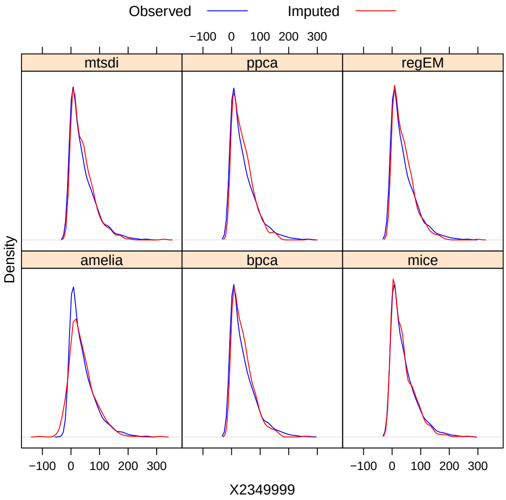


Figure 16. Kernel density estimates for the marginal distributions of the observed and imputed values at station X2349999 under scenario 7 - (decendial rainfall and removing 20 periods of 12 months).

Preliminary results

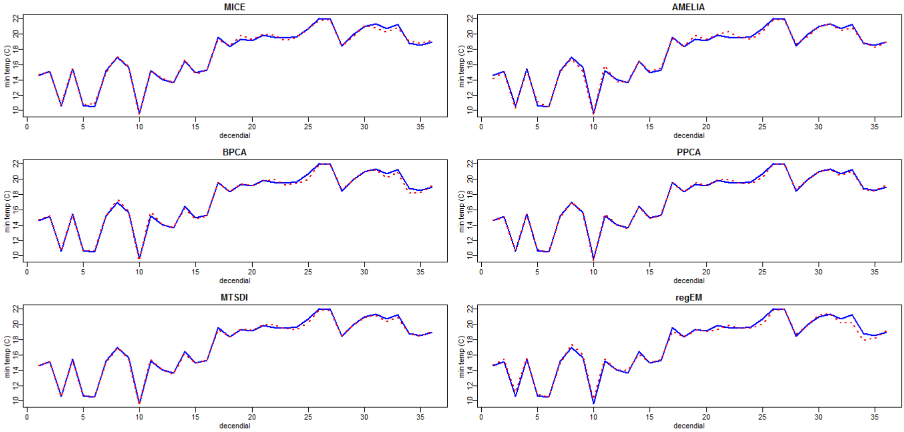
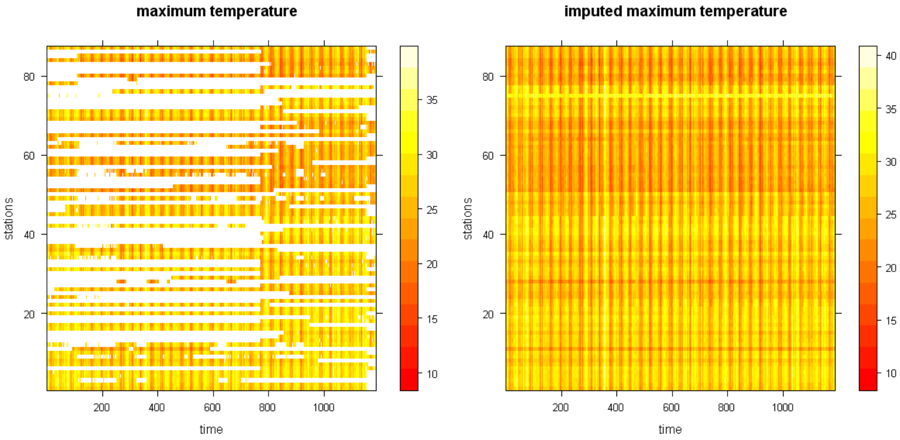


Figure 17. Direct comparison between decendial estimates (red dashed lines) and observed data (blue solid lines) for the six imputation methods at station X2349999 (first subset) under scenario xx - (decendial minimum temperature and removing 20 periods of 1 year).

Preliminary results



Preliminary results

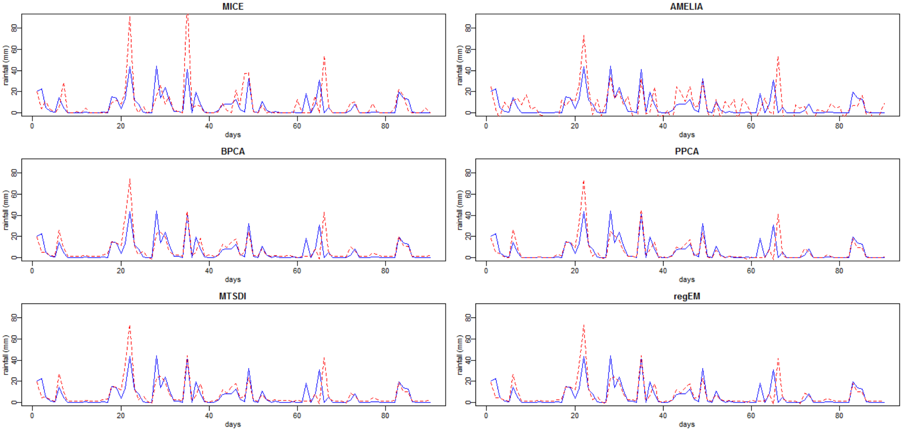


Figure 19. Direct comparison between daily estimates (red dashed lines) and observed data (blue solid lines) for the six imputation methods at station X2349999 (first subset) under scenario xx - (daily rainfall and removing 20 periods of 3 months).

Simolo's approach (2009)

X[i,j] - precipitação dia i, estação j
N - número de dias
n.c[j] - número dias com chuvas estação j
Para cada estacao
Para cada dia com dado > 0

- i - Procure n1=150 dados positivos mais “próximos” no tempo
- ii - Estime parâmetros da densidade gamma(a,b)
- iii - Calcule p1=p(X[i,j]<x[i,j]/a,b)
- i2 - Procure n2=1000 dados positivos mais ”próximos”no tempo
- ii2 - Estime parâmetros da densidade gamma(a2,b2)
- iii2 - Calcule p2=p(X[i,j]<x[i,j]/a2,b2)

Para cada estação
Para cada dia sem dado

- Calcule p1.hat (media ponderada dos p1 vizinhos no espaço)
- Calcule p2.limiar, tal que p2.vizinhos = n.c.vizinhos/N
- Faça C[i,j] = 1 se p1.hat > p2.vizinhos
- Se C[i,j] = 1, estime x[i,j]

Future work

- to standardize a methodology to check the consistency of weather data;
- sensitivity analysis varying the dimensionality of the problem and the proportion of missing values;
- implications of improper imputations on pricing crop insurance contracts;
- better methods to impute daily rainfall;
- to evaluate the accuracy of different interpolation methods for weather variables;
- toolkit with imputation and comparison methods available as an R package.

Future work

- to standardize a methodology to check the consistency of weather data;
- sensitivity analysis varying the dimensionality of the problem and the proportion of missing values;
- implications of improper imputations on pricing crop insurance contracts;
- better methods to impute daily rainfall;
- to evaluate the accuracy of different interpolation methods for weather variables;
- toolkit with imputation and comparison methods available as an R package.

Future work

- to standardize a methodology to check the consistency of weather data;
- sensitivity analysis varying the dimensionality of the problem and the proportion of missing values;
- implications of improper imputations on pricing crop insurance contracts;
- better methods to impute daily rainfall;
- to evaluate the accuracy of different interpolation methods for weather variables;
- toolkit with imputation and comparison methods available as an R package.

Future work

- to standardize a methodology to check the consistency of weather data;
- sensitivity analysis varying the dimensionality of the problem and the proportion of missing values;
- implications of improper imputations on pricing crop insurance contracts;
- better methods to impute daily rainfall;
- to evaluate the accuracy of different interpolation methods for weather variables;
- toolkit with imputation and comparison methods available as an R package.

Future work

- ➢ to standardize a methodology to check the consistency of weather data;
- ➢ sensitivity analysis varying the dimensionality of the problem and the proportion of missing values;
- ➢ implications of improper imputations on pricing crop insurance contracts;
- ➢ better methods to impute daily rainfall;
- ➢ to evaluate the accuracy of different interpolation methods for weather variables;
- ➢ toolkit with imputation and comparison methods available as an R package.

Future work

- to standardize a methodology to check the consistency of weather data;
- sensitivity analysis varying the dimensionality of the problem and the proportion of missing values;
- implications of improper imputations on pricing crop insurance contracts;
- better methods to impute daily rainfall;
- to evaluate the accuracy of different interpolation methods for weather variables;
- toolkit with imputation and comparison methods available as an R package.

Some References

Abayomi, K., Gelman, A. and Levy, M. (2008) Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society Series C, Applied Statistics*, **57**, 273–291.

Honaker, J., King, G. and Blackwell, M. (2009) AMELIA II: A Program for Missing Data. `http://gking.harvard.edu/amelia`

Junger, W.L., Ponce de Leon, A. and Santos, N. (2003) Missing data imputation in multivariate time series via EM algorithm. *Cadernos do IME*, **15**, 8–21.

Oba, S., Sato, M. Takemasa, I., Monden, M. Matsubara, K. and Ishii, S. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Schneider, T. (2001) Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. *Journal of Climate*, **14**, 853–871.

Stacklies, W., Redestig, H., Scholz, M., Walther, D. and Selbig, J. (2007) pcaMethods-a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164–1167.

Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. (2006) Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, **76**, 1049–1064.