Regressão simbólica sobre séries temporais de dados meteorológicos utilizando programação genética

Programa de Pós-Graduação em Computação Aplicada
Departamento de Informática
Universidade Tecnológica Federal do Paraná

Roberto Santos - roberto@simepar.br
Heitor Lopes - hslopes@utfpr.edu.br

24 de outubro de 2012

Introdução

Contexto

Este trabalho apresenta um método para modelagem e preenchimento de falhas em séries temporais. O método é baseado em programação genética, uma técnica de computação evolucionária e são utilizados os dados de temperatura média da estação meteorológica de Palotina/PR.

Dados meteorológicos

WMO (2006)

"Os seres humanos vivem em um ambiente cercado pela atmosfera. Desta forma, todas as mudanças e fenômenos que ocorrem na atmosfera e no ambiente onde vive-se afetam direta ou indiretamente os seres humanos. Ser capaz de minimizar os efeitos negativos destes fenômenos e usar os resultados de forma benéfica para os seres humanos é uma das motivações para observar a atmosfera e o meio ambiente."

Estação meteorológica automática

É formada por um conjunto de sensores responsáveis por observar as mudanças e fenômenos da atmosfera, armazenando e enviando estes dados periodicamente. Os sensores mais comuns registram dados de temperatura, pressão, radiação solar, direção/velocidade do vento e precipitação acumulada.

Problema

Ausência de dados

Por diversas razões é comum a ausência de dados em séries temporais de dados meteorológicos. Estas ausências estão associadas à:

- Falha no sensor
- Falha na transmissão do dado
- Reprovação do dado por um sistema de controle de qualidade

Implicações

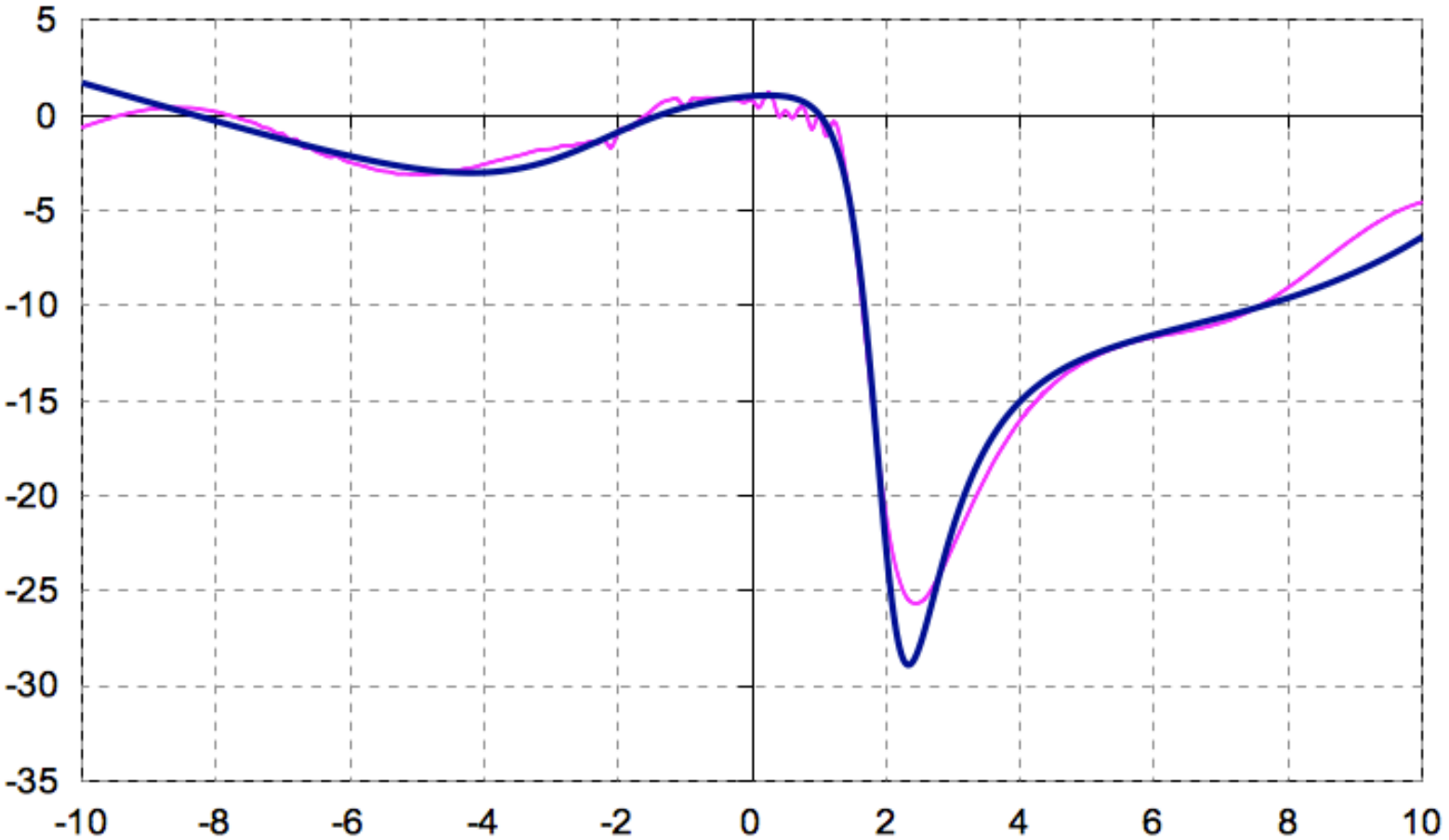
Diversos modelos utilizados na agricultura (balanço hídrico) utilizam os dados meteorológicos sequencialmente e a ausência de dados impede que o modelo seja executado.

Regressão simbólica

A regressão simbólica consiste em induzir expressões matemáticas a partir de dados de um sistema matemático através da manipulação de expressões. O sistema matemático é descrito por dados de entrada e saída, que consistem em valores de uma função desconhecida f, tal que $f : R_n \rightarrow R$, ou seja, casos de fitness geralmente têm um formato:
$x_1, x_2, ..., x_n, y$
onde x_1 a x_n representam as variáveis independentes no sistema e y, a variável dependente.

Entrada	Saída
...	...
-3.9	-2.9972
-3.6	-2.8780
-3.5	-2.8184
-3.4	-2.7487
...	...

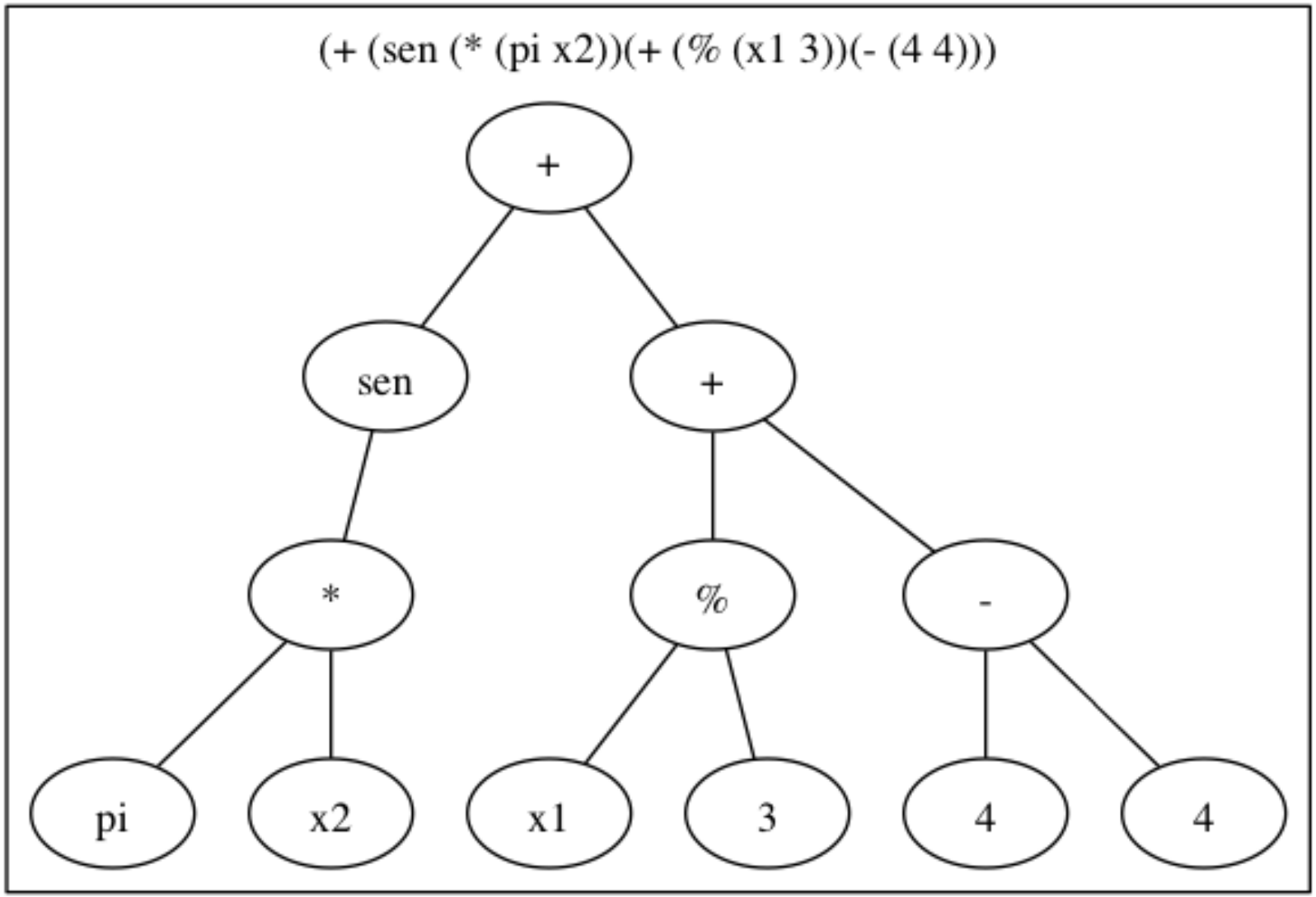
Regressão simbólica



Programação genética (PG)

- Os fundamentos da PG surgiram a partir da tese de doutorado de John Koza (1972), orientada por John Holland
- A PG é um descendente direto do AG
- Baseia-se no conceito de "construção de programas" para resolver problemas
- Programas = funções + terminais
- O espaço de busca de todos os programas possíveis é infinito e intratável

Programação genética (PG)

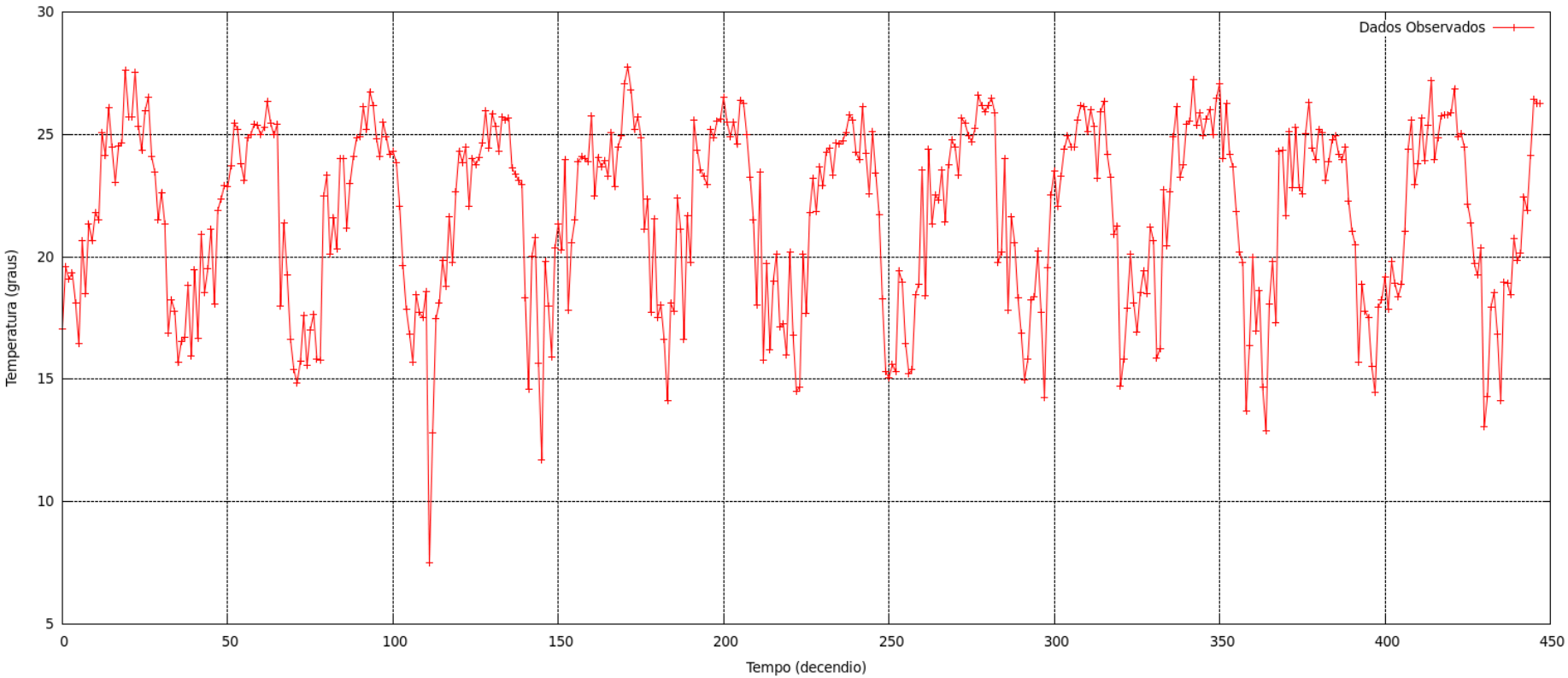


Passos preparatórios para PG

1. Conjunto de terminais
2. Conjunto de funções
3. Casos de *fitness*
4. Medidas de *fitness*
5. Parâmetros de controle e variáveis qualitativas
6. Critério de parada
7. Especificação do resultado

Tratamento dos dados - Casos de *fitness*

- Estação meteorológica de Palotina/PR
- Variável: temperatura média
- Período: jul/1997 a mar/2011
- Agrupamento dos dados em períodos de 10 dias (decêndios)
- 449 casos de *fitness*



Definição dos parâmetros

- Funções: $+, -, *, /, \mathit{seno}, \mathit{sigmoide}, \mathit{sinc}$
- Terminais: $[0, 1], 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, \pi, x, x_{-1}, x_{-2}, x_{-3}$
- Medida de *fitness*: $f = \sum_{i=1}^{n} |x_i - y_i|$
- Geração da população inicial: ramped half-and-half
- Seleção: torneio de tamanho 7
- Número de populações: 1
- Tamanho da população: 50.000
- Operadores genéticos: reprodução e recombinação
- Profundidade da árvore: 10
- Probabilidade de recombinação: 0,9
- Probabilidade de reprodução: 0.1
- Critério de parada: $100 + 1$

Resultados preliminares

Representação S

(* π (+ (+ (+ (- x x3) (sigm (sinc (+ (* x 0.056) (* π 0.971))))) (- x x3)) (sigm (* (* (+ (+ π (+ π π)) (+ (* (* x 0.056) 0.056) (* (- x x3) 0.971))) (sigm (* 0.537) (sinc (* 0.054 x))))) (sinc (- (sigm (+ (- x x3) (- x x3))) (+ (+ π π) (* x 0.056)))))))

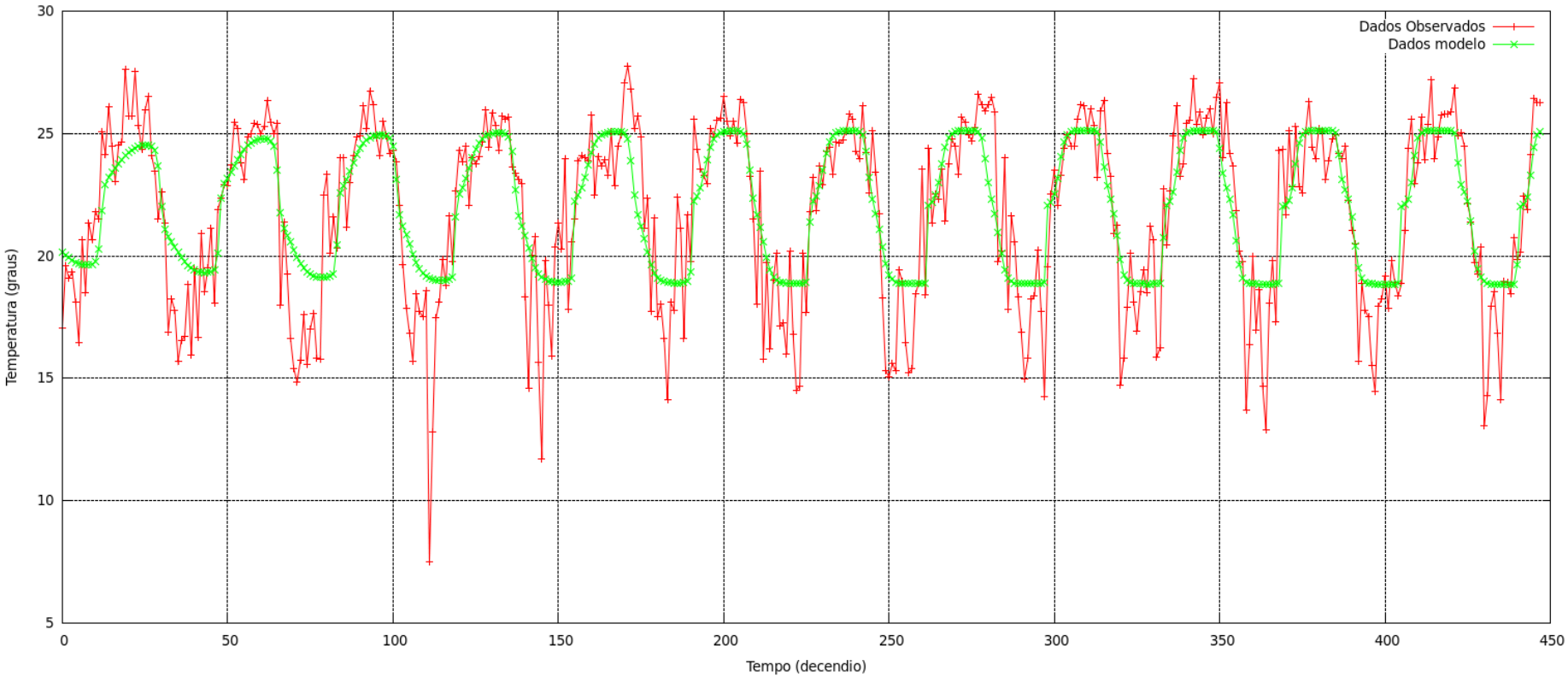
Coeficiente de Pearson (R)

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, [-1, +1] \quad (1)$$

$$R = 0.83 \quad (2)$$

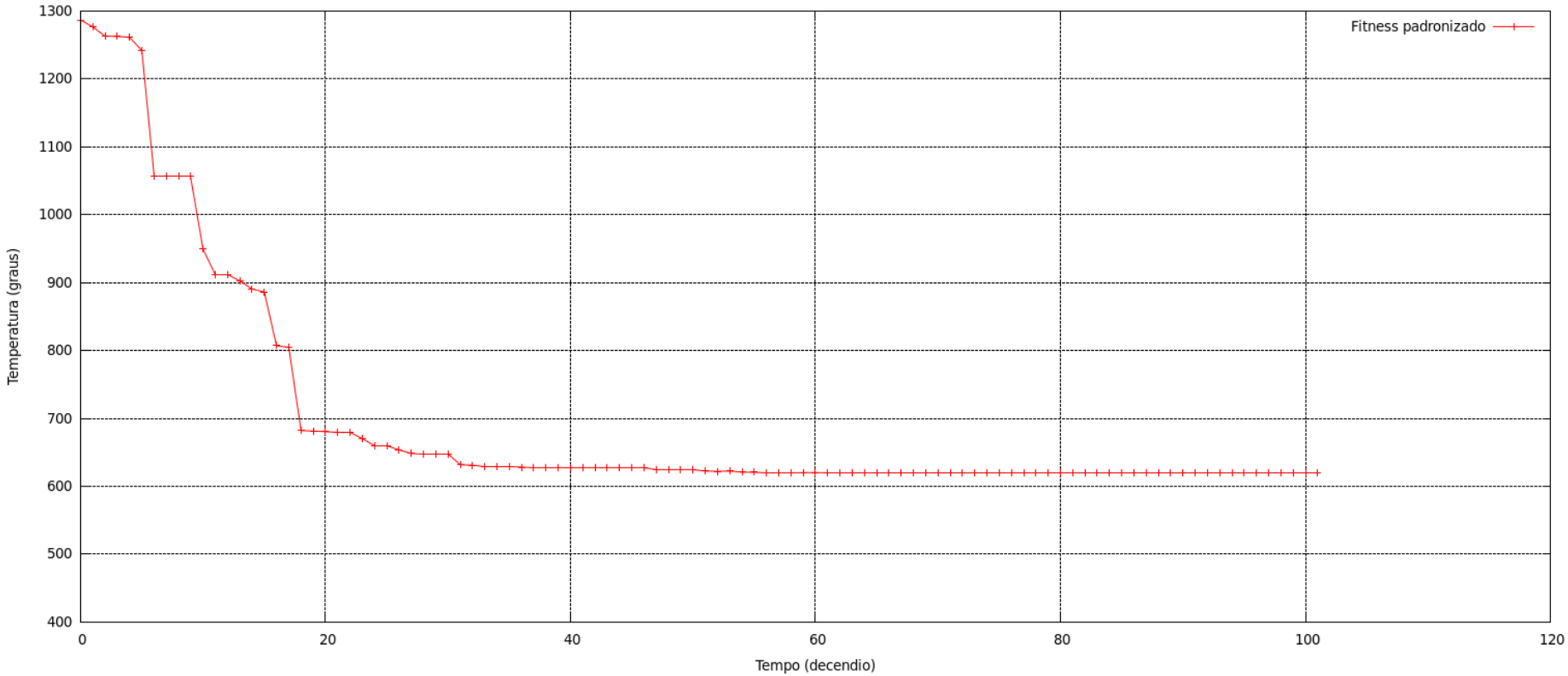
Resultados preliminares

Dados observados x Modelo PG



Resultados preliminares

Fitness cru



Conclusões

Pontos observados

- Comportamento aproximado da realidade
- Facilidade de uso do método (ecj)
- Necessidade de limitar a profundidade da árvore da solução
- Necessidade de identificar as funções adequadas ao fenômeno físico associado

Próximos passos

- Comparar os resultados utilizando Programação de Expressão Genética
- Utilizar somente dados aprovados no controle de qualidade
- Utilizar períodos menores (1 ano) e/ou a média de cada decêndio

Conclusões

Agradecimentos

Os autores agradecem ao SIMEPAR pelo fornecimento dos dados utilizados neste trabalho.