Management and Analysis of Spatial-Time Series

1º SciDisc Workshop

Eduardo Ogasawara
http://eic.cefet-rj.br/~eogasawara

Inria informatics mathematics

COPPE UFRJ
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

National Laboratory Scientific Computing

Instituto de computação ufrj

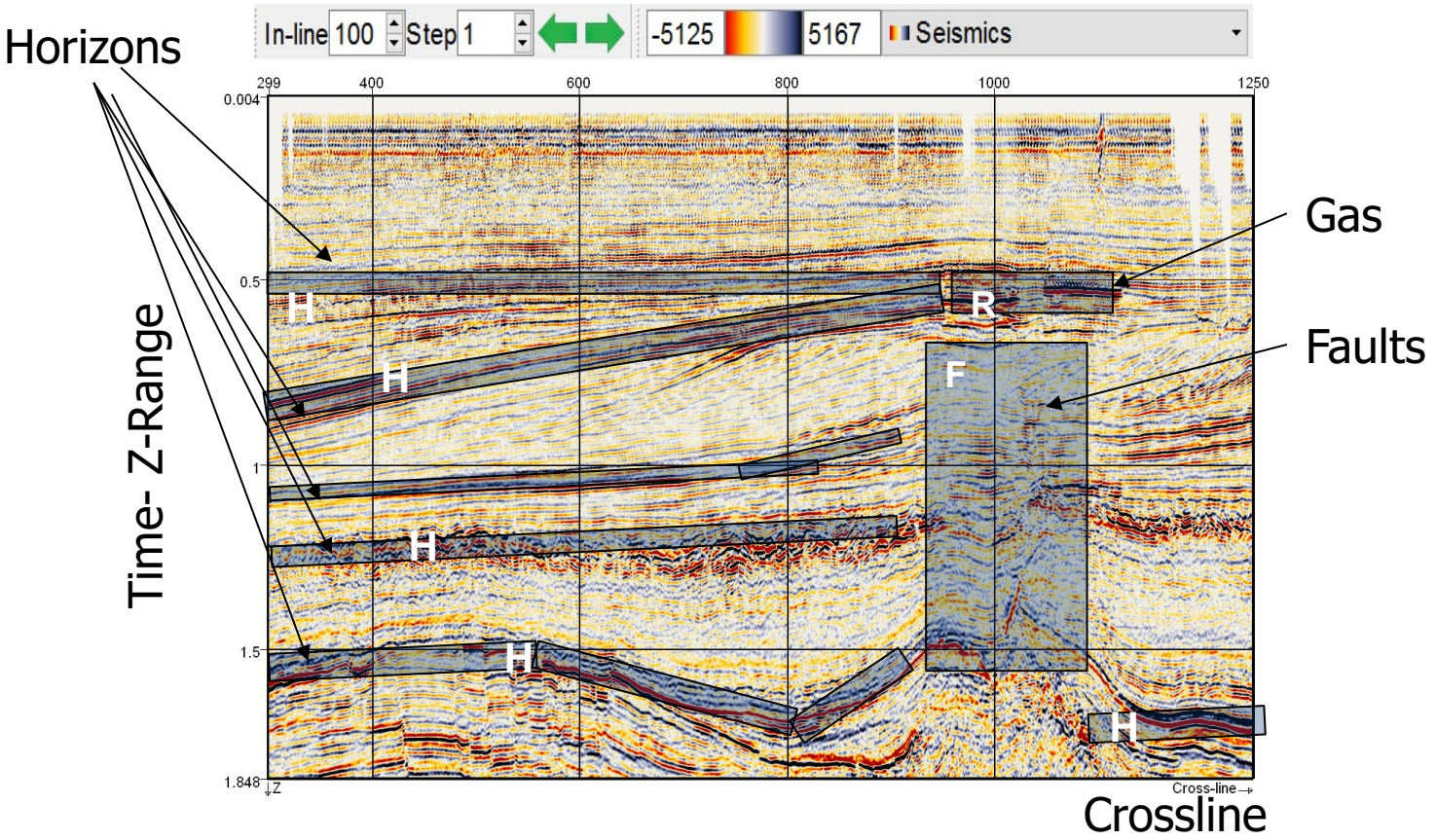
CEFET/RJ

Knowledge Discovery in Big Data

- Data Deluge
  - Science: astronomy, Seismic
  - Business/Persons: IoT, Flights
  - Government: Smart cities Urban mobility
- Challenges for Knowledge Discovery
  - Data management
  - Data analysis
    - Prediction, Classification e Pattern Identification
- Many phenomena are modeled in space-time

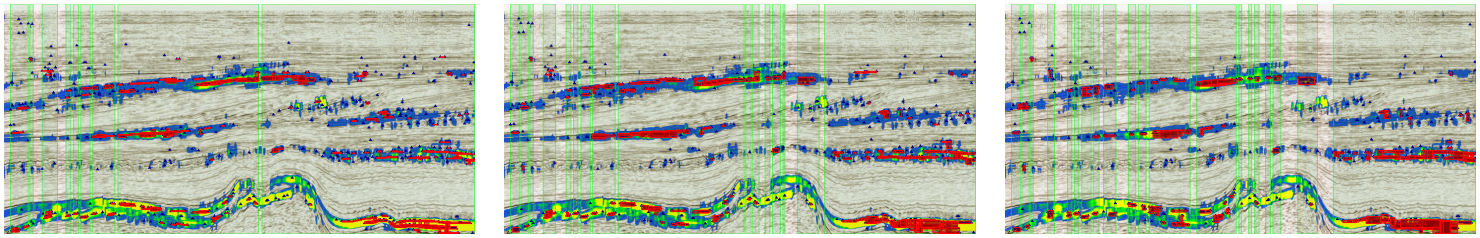
Seismic Analysis

- 2D Slice of seismic dataset (inline 100)



Seismic Analysis – Results

- Motifs Analysis
  - Discovering spatial-time motifs in seismic datasets
  - Murillo Dutra master degree
- Sequence Mining of Spatial-Time Series
  - Identification of solid spatial-time sequences
  - Riccardo Campisano master desgree



$<a, j>$ $(\mathbf{a})$ $<a, a, j>$ $(\mathbf{b})$ $<a, a, j, j>$ $(\mathbf{c})$

Fig. 4. Potential *bright spots* identified using the proposed algorithm for *inline* 401, alphabet size 10, γ = 80%, and δ = 20%. The results follow the blue-yellow pattern produced using the previously known *bright spots* for this dataset. [5].

Seismic Analysis – Research Opportunities

- 3D Analysis (x, y, and time)
  - Solid Cube Patterns
- Techniques for faults detection
  - Intuition that absence of solid patterns drives faults detection
- Techniques for shape detections
  - Combinations of motifs/solid patterns
- Comparison between motifs identification and sequence mining

Flight Delays

Brazilian Flights Dataset
Airports Meteorological Dataset

Flight Delays – Results

- Data warehouse
  - Brazilian National Flights
  - Meteorological condition

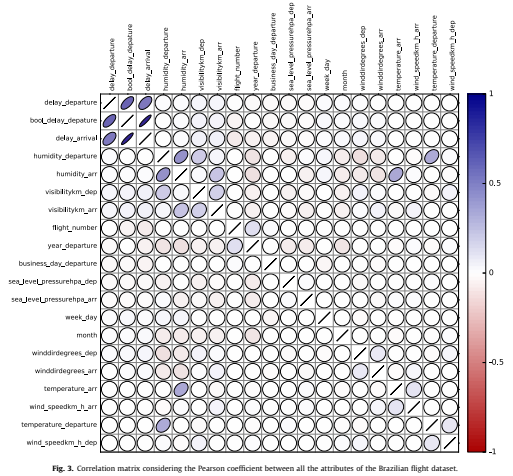


Fig. 3. Correlation matrix considering the Pearson coefficient between all the attributes of the Brazilian flight dataset.

- Identification of frequent patterns that leads to delays
  - Alice Sternberg master degree

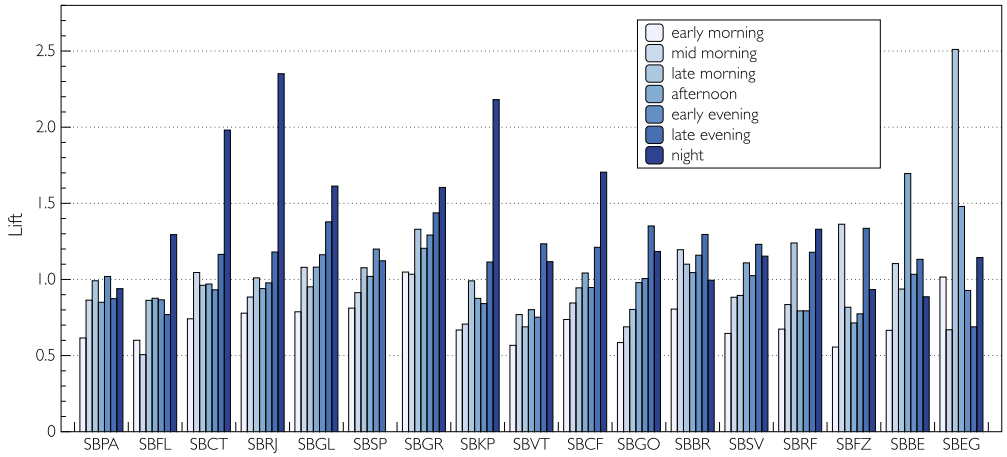


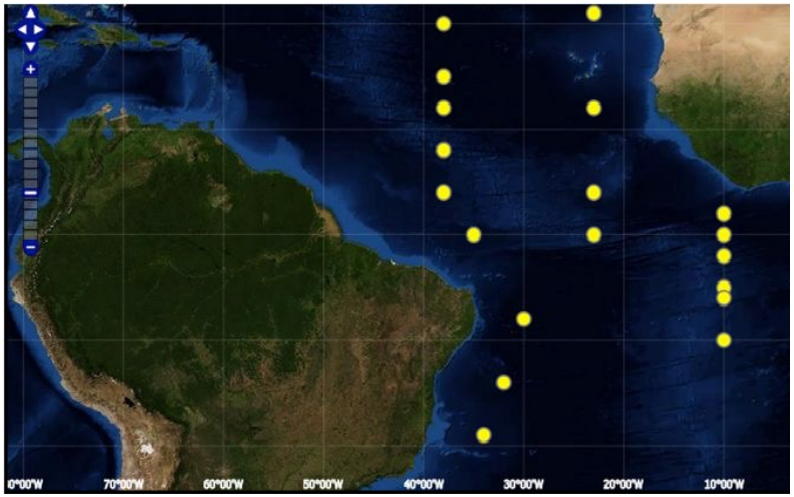
Fig. 7. Lift analysis of the rules containing the airport and the time of departure on the antecedent and a delay on the consequent – the airports are ordered from south to north.

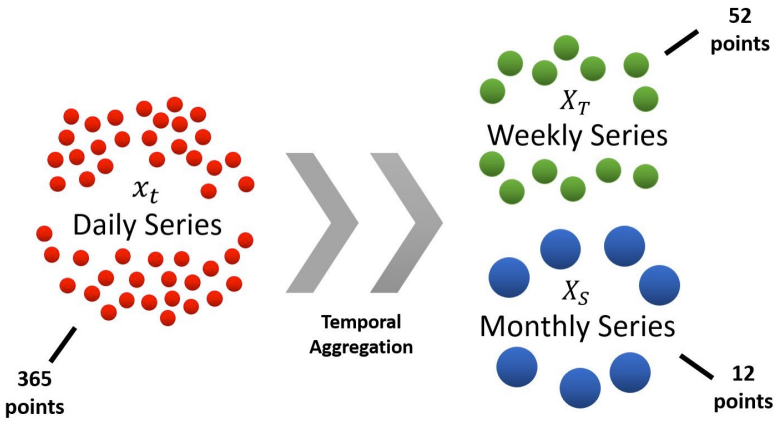
Flight Delays – Research Opportunities

- Airport delays propagation
  - On going
- Flight delays propagation
  - On going
- Prediction of flight delays
  - On going*
- Replication of techniques using American datasets

Time-Series Prediction

- Long term prediction of sea surface temperature





Time-Series Prediction – Results

- Framework for analysis of prediction performance compared to linear models

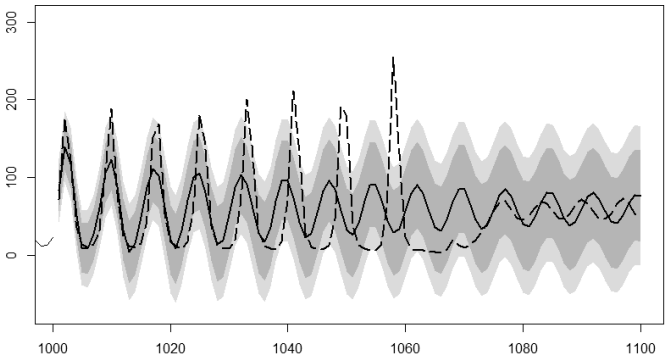


Fig. 2: ARMA predictions (solid line) for the time series A of the Santa Fe Competition. The actual time series values are represented by the dashed line.

TABLE III: Rankings of the top 25 results of the chosen competition datasets including results from TSPred R-package

Rank	Santa Fe				EUNITE		CATS			NN3		NN5	
	Dataset A		Dataset D							Dataset A		Dataset A	
	index	NMSE	index	NMSE[1]	Participant	MAPE [%]	Participant	E1	E2	Participant	Mean SMAPE	Participant	Mean SMAPE
1	W	0.02	ZH	0.08	Chih-Jen Lin	1.982	Sarkka*	408	346	Illies*	15.18%	Andrawis	20.40%
2	Sa	0.08	**TSPred$_{(ARIMA)}$**	**0.54**	Esp	2.149	Cai*	441	402	Adeodato*	16.17%	Vogel	20.50%
3	M	0.38	U	1.30	Brockmann	2.498	Kurogi*	502	418	Flores*	16.31%	D'yakonov	20.60%
4	L	0.45	**TSPred$_{(PR)}$**	**1.61**	**TSPred$_{(PR)}$**	**2.779**	Hu*	530	370	Chen*	16.55%	Rauch	21.70%
5	U	0.62	Z	4.80	Zivcak	2.873	Palacios-Gonzalez	577	395	D'yakonov	16.57%	Luna	21.80%
6	A	0.71	C	6.40	Kowalczyk	2.985	Maldonado*	644	542	Kamel*	16.92%	Wichard	22.10%
7	McL	0.77	W	7.10	Lewandowski	3.223	Simon*	653	351	Abou-Nasr	17.54%	Gao	22.30%
8	**TSPred$_{(ARIMA)}$**	**0.90**	S	17.00	Kowalczyk	3.264	Verdes*	660	442	Theodosiou*	17.55%	Puma-Villanueva	23.70%
9	**TSPred$_{(PR)}$**	**0.99**			Ortega	3.380	Chan*	676	677	**TSPred$_{(ARIMA)}$**	**17.79%**	Dang	25.30%
10	N	1.00			King	3.388	Wichard*	725	222	de Vos	18.24%	Pasero	25.30%
11	P	1.30			Lotfi	3.389	Beliaev*	928	762	Yan	18.58%	Adeodato	25.30%
12	Can	1.40			Guijarro	3.421	Kong	954	994	C49	18.72%	undisclosed	26.80%
13	K	1.50			Weizenegger	3.694	Wang	1037	402	Perfilieva*	18.81%	undisclosed	27.30%
14	Sw	1.50			**TSPred$_{(ARIMA)}$**	**3.820**	Cellier*	1050	278	Kurogi*	19.00%	**TSPred$_{(ARIMA)}$**	**27.80%**
15	Y	1.50			Boger	3.958	Crone*	1156	995	Beadle	19.14%	Tung	28.10%
16	Car	1.90			Bontempi	3.997	**TSPred$_{(ARIMA)}$**	**1173**	**917**	Lewicke	19.17%	undisclosed	33.10%
17					Pelikan	4.348	Acernese*	1247	1229	Sorjamaa*	19.60%	undisclosed	36.30%
18					Brockmann	4.373	Yen-Ping*	1425	894	Isa	20.00%	undisclosed	41.30%
19					Pelikan	4.437	**TSPred$_{(PR)}$**	**7387**	**6778**	C28	20.54%	**TSPred$_{(PR)}$**	**41.50%**
20					Rivieccio	4.502				Duclos-Gosselin	20.85%	undisclosed	45.40%
21					Brockmann	4.580				Papadaki*	22.70%	undisclosed	53.50%
22					Ivakhnenko	4.653				Hazarika	23.72%		
23					Brockmann	4.712				C17	24.09%		
24					Brockmann	5.087				Njimi*	24.90%		
25					Brockmann	5.425				Pucheta*	25.13%		

* et al.
[1] NMSE error for the 15 first predicted observations

Time-Series Prediction – Results

- Effect of temporal aggregation for long-term prediction of sea surface temperature
  - Scientific Initiation of Rebecca Salles

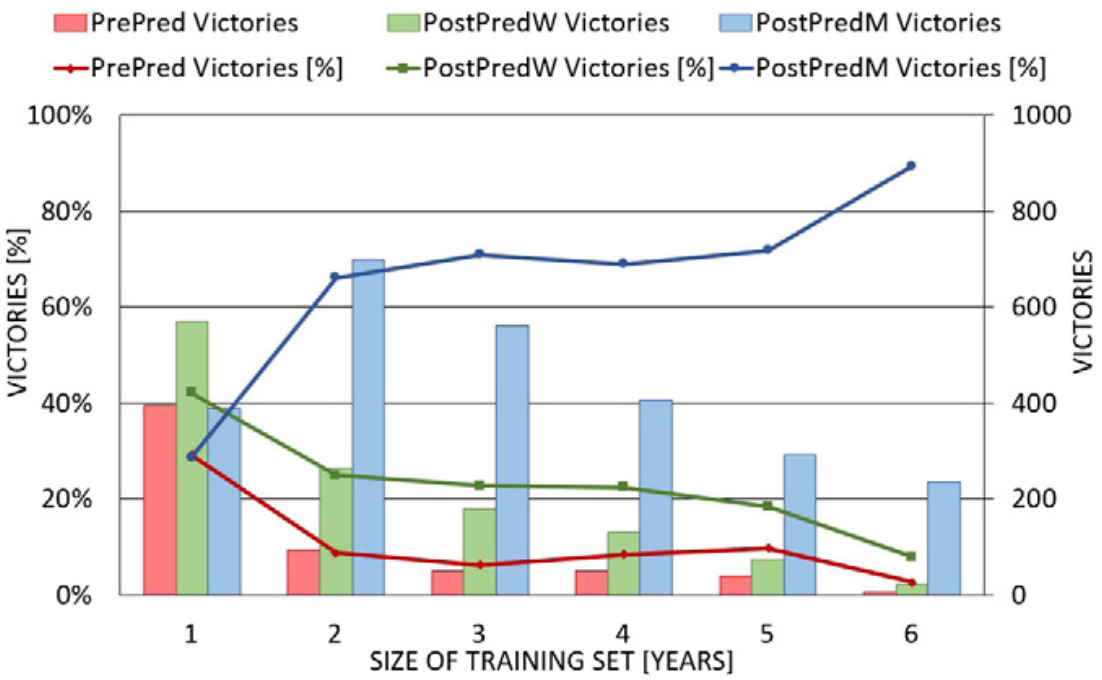


Fig. 8. Graphic of the victories of each prediction approach regarding their performances in generating up to twelve monthly aggregated forecasts.

Time-Series Prediction – Research Opportunities

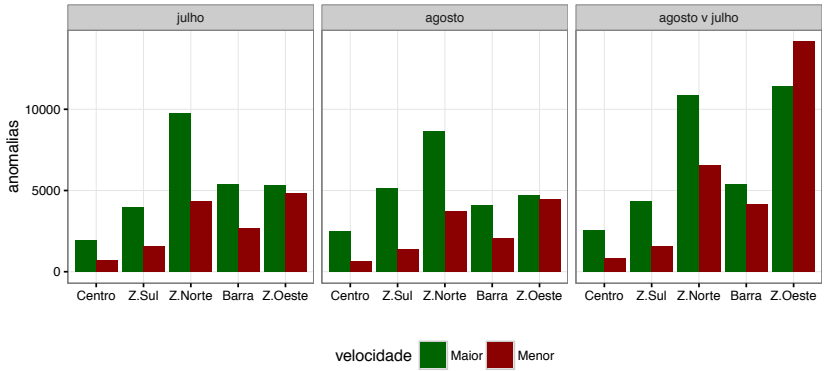
- Expansion of framework prediction for machine learning methods
  - On going
- Study of different preprocessing methods for supporting non-stationarity
  - On going
- Creation of novel methods for non-stationarity for machine learning methods

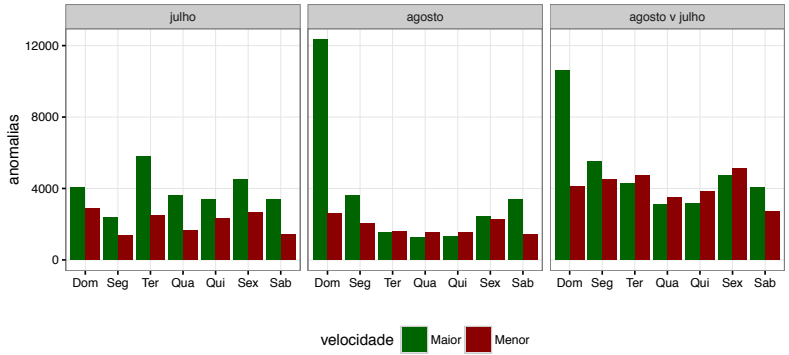
Urban Mobility

Approximately more than 4 million of observations per day
Bus as trajectory sensors
Spatial-Temporal Aggregation: Regions as virtual sensors

Urban Mobility – Results

- Data collection (done by UFF)
- Data Cleaning, Spatial-Time Aggregation
- Preliminary Analysis of Anomalies
  - Ana Beatriz Cruz Master Degree Theme





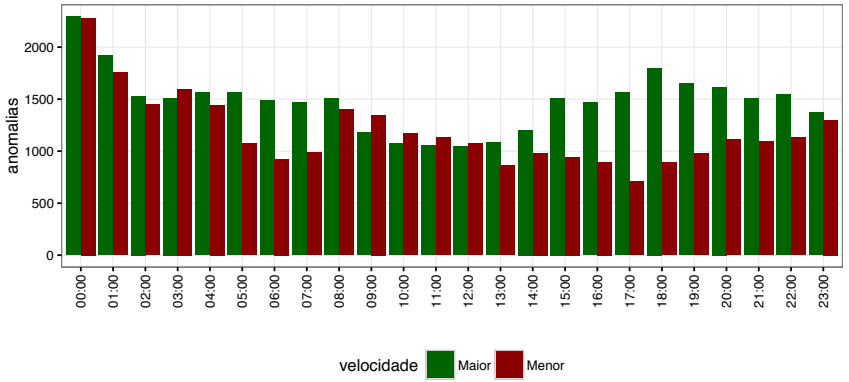


Figura 3. Anomalias identificadas por faixa de horário (ago v julho)

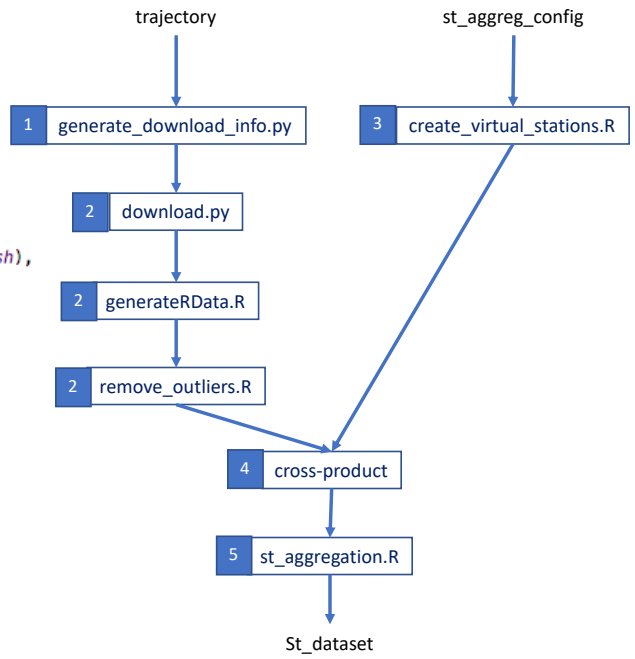
Urban Mobility – Research Opportunities

- Persistence and Querying
- Trajectory or Aggregated analysis
- Identification of Patterns, Anomalies, and Paradigm

Parallel and Distributed Execution Using Spark

- João Ferreira Master Degree Theme

```
1  val trajectory: Relation = Relation(Schema(key, initialTime, endTime),
2    Tuple("copa-do-mundo-2014", "2014-06-01", "2014-07-31"))
3  val st_aggreg_config: Relation = Relation(Schema(radius, interval, busesMesh),
4    Tuple("10", "10", "malha-2014.csv"))
5  w = Workflow("2014CupAggregation", () => {
6    r1 = SplitMap(Activity("generate_download_info.py"), key, trajectory)
7    r2 = Map(Activity("download.py"), r1)
8    r3 = Map(Activity("generateRdata.R"), r2)
9    r4 = Map(Activity("remove_outliers.R"), r3)
10   r5 = Map(Activity("create_virtual_stations.R"), st_aggreg_config)
11   r6 = Query(CrossProduct, r4, r5)
12   result = Map(Activity("st_aggregation.R"), r6)
13 })
14 w.execute()
```



(a) (b)

Figura 2. Workflow para análise de tráfego durante a COPA de 2014 : a) Especificação do Workflow usando linguagem Scala; b) grafo mostrando as dependências entre as atividades

Recent Published Papers Related to the Project

- Ferreira J. et al, 2017 - Uma Proposta de Implementação de Álgebra de Workflows em Apache Spark no Apoio a Processos de Análise de Dados. In: BreSci
- Salles R. et al. 2017 - A Framework for Benchmarking Machine Learning Methods Using Linear Models for Univariate Time Series Prediction, IJCNN
- Marinho A. et al. 2017 - Deriving scientific workflows from algebraic experiment lines: A practical approach. Future Generation Computer Systems.
- Guedes G. et al. 2016 - Discovering top-k Non-Redundant Clusterings in Attributed Graphs. Neurocomputing.
- Sternberg A. et al., 2016 - An analysis of Brazilian flight delays based on frequent patterns. Transportation Research. Part E, Logistics and Transportation Review
- Salles R. et al, 2016 - Evaluating Temporal Aggregation for Predicting the Sea Surface Temperature of the Atlantic Ocean. Ecological Informatics.
- Machado E. et al, 2016 - Exploring machine learning methods for the Star/Galaxy Separation Problem. In: IJCNN
- Cruz A. et al, 2016 - Identificação de Motifs em Agregações de Séries Espaço-Temporais de Mobilidade Urbana. In: WTDBD/SBBD
- Campisano, R., Porto. F., Pacitti, E., Florent M., Ogasawara E., Spatial Sequential Pattern Mining for Seismic Data. In: SBBD
- Salles et al., 2015 - Evaluating Linear Models as a Baseline for Time Series Imputation. In: SBBD
- …
- Ogasawara, E. et al., 2010 Adaptive Normalization: A Novel Data Normalization Approach for Non-Stationary Time Series. In: IJCNN.

CEFET/RJ Team