



# Online data mining services for dynamic spatial databases

---

Manuel Costa, Inês Sousa, Alexandra Fonseca, Luís Teixeira, Alvaro Barbosa, Jorge Cardoso, Vasco Carvalho, Ivan Franco, Nuno Capeta, Marco Orlando, Paulo Castro, Diana Henriques and Paulo Rosa



# Motivation

**Large databases updated by automatic processes.**

**There aren't many commercial solutions to execute data mining algorithms at server side.**

Automatic processes to analyze dynamic data represent an appealing market (e.g. **traffic, air quality, water quality**).



# Goals

Develop a **Service Oriented Architecture (SOA)** with data mining functionalities

Direct access to dynamic spatial Data Bases and Data Mining algorithms

Develop customized applications



# Outline

Brief overview of Data Driven Modeling

Data mining services in action : SNIRH Data Mining

Technical description of the services

Case Studies : Spatial sound data mining



# Data Driven Modeling

Robust methods to model relations between data (e.g. Artificial Neural Networks ANN).

Modeling with insufficient knowledge about complex systems => 100% based on data: “you don’t have to know the underlying physics”.



# Data Driven Modeling Limitations

Modeling often requires computational muscle.

Some methods do not scale for large volumes of data.



www.ydreams.com

# SNIRH Data Mining



**SNIRHParameters - Microsoft Internet Explorer**

File Edit View Favorites Tools Help

Address: <http://qdm.inag.pt/datamining/Snrhdbrowser/SNIRHSelection.aspx#>

Google - Search Web PageRank 54 blocked Autofill Options

Logged by manel Logoff

INSTITUTO DA ÁGUA

SNIRH Data Mining GRID Computing

Home Nova Tarefa Estatísticas-Resultados ANN-Resultados

Submeter Tarefa - Passo 1 de 4 : Selecionar as Estações

Avançar

Redes de monitorização

Hidrométrica automática

Bacias

TEJO

Estações disponíveis

Estações selecionadas

Adicionar

Estações disponibilizadas

Estações selecionadas

Done

Map showing the Tejo basin with monitoring stations highlighted in green. A legend on the right lists various monitoring networks and auxiliary layers.

**ANNTaskResult - Microsoft Internet Explorer**

File Edit View Favorites Tools Help

Address: <http://qdm.inag.pt/datamining/Snrhdbrowser/ANNTaskResult.aspx?taskId=1a9441d2-d2a3-44f3-ae98-b2b4551b8d6>

Google - Search Web PageRank 54 blocked Autofill Options

Logged by manel Logoff

INSTITUTO DA ÁGUA

SNIRH Data Mining GRID Computing

Home Nova Tarefa Estatísticas-Nova Tarefa Estatísticas-Resultados ANN-Nova Tarefa ANN-Resultados

Estatísticas Esp./Obs. Código em VBA Gráfico Download dos Dados

Gráfico

Dependente : 15G/02 : AGR0AL : Nível hidrométrico Instantâneo Percentil\_95  
Esperada : 15G/02 : AGR0AL : Nível hidrométrico Instantâneo Percentil\_95

Tarefas Submetidas

Almourol Hidro dia - P05, P50, P95 - 8, 250000  
Agrual Hidro a 3 dias - 6, 250000  
Tomar  
Monchique Novo OD Médio a 3 dias - 8, 250000

Parâmetros da tarefa

Nome da tarefa: Agrual Hidro a 3 dias - 8, 250000  
N. de neurónios: 6  
N. de neurónios na "Hidden Layer": 8  
N. de iterações: 250000  
Ratio de dados para treino: 90,0  
Data inicio: 22-09-2001 0:00:00  
Data fim: 26-04-2005 0:00:00  
Intervalo de agregação: 1 day

Variáveis independentes

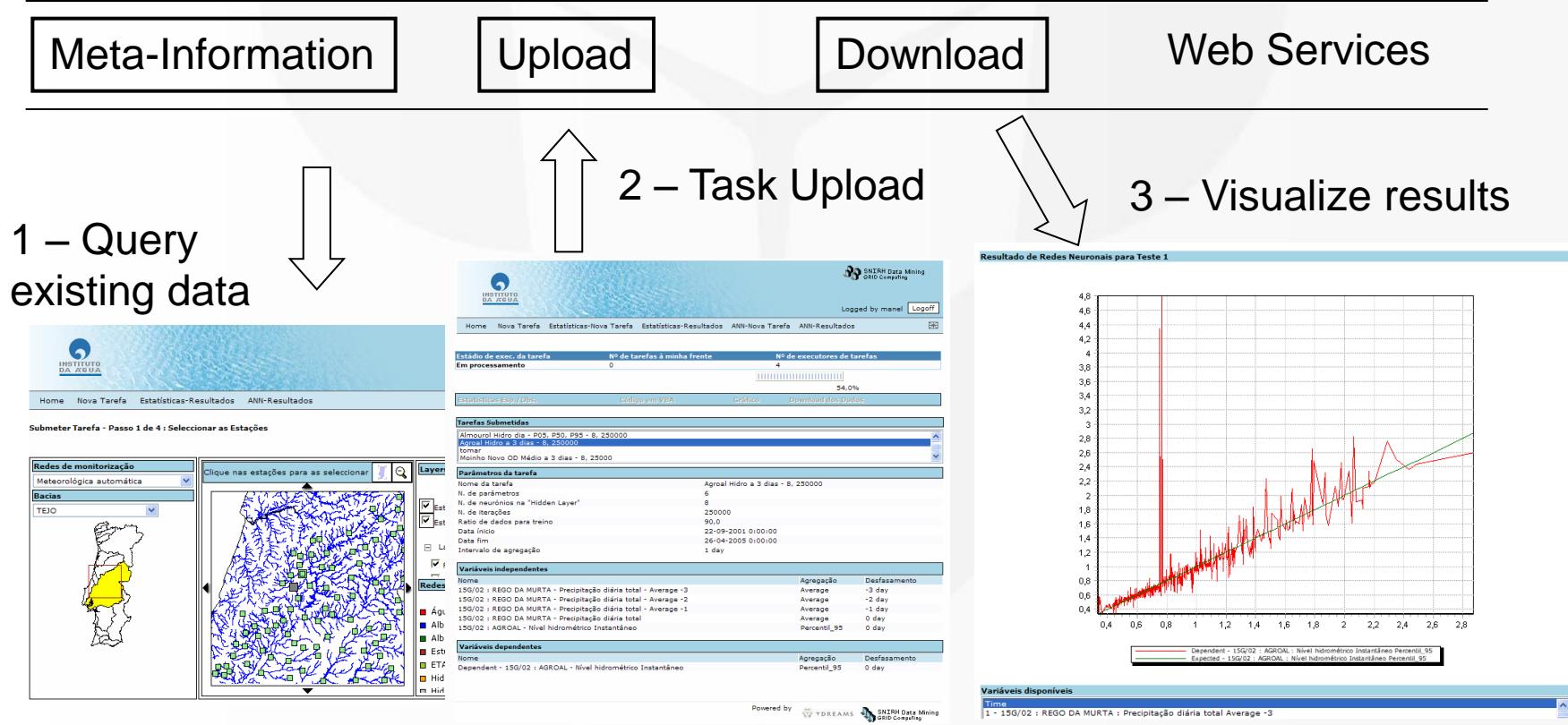
Nome	Agregação	Desfasamento
15G/02 : REGO DA MURTA - Precipitação diária total - Average -3	Average	-3 day
15G/02 : REGO DA MURTA - Precipitação diária total - Average -2	Average	-2 day
15G/02 : REGO DA MURTA - Precipitação diária total - Average -1	Average	-1 day
15G/02 : REGO DA MURTA - Precipitação diária total	Average	0 day
15G/02 : AGR0AL - Nível hidrométrico Instantâneo Percentil_95	Percentil_95	0 day

Variáveis dependentes



www.ydreams.com

# Online Data Mining Service (User's perspective)





# Online Data Mining Service (Server's perspective)

Meta-Information

Upload

Download

Web Services



2 – Task Upload

The screenshot shows the SHIRI Data Mining interface with the following details:

**Logged in user:** manel

**Task Status:** Em processamento (Processing) - 0 tasks, 4 executors, 54.0% complete.

**Submitted Tasks:**

- Almourol Hidro dia - P05, P50, P95 - 8, 250000
- Almourol Hidro a 3 dias - 8, 250000
- tomar
- Moinho Novo OD Médio a 3 dias - 8, 25000

**Task Parameters:**

Nome da tarefa	Agroal Hidro a 3 dias - 8, 250000
N. de tarefas	8
N. de neurônios na "Hidden Layer"	8
N. de iterações	250000
Ratio de dados para treino	90.0
Data inicio	22-09-2001 0:00:00
Data fim	26-04-2005 0:00:00
Intervalo de agregação	1 day

**Independent Variables:**

Nome	Agregado	Desfasamento
150/02 / REGO DA MURTA - Precipitação diária total - Average -3	Average	-3 day
150/02 / REGO DA MURTA - Precipitação diária total - Average -2	Average	-2 day
150/02 / REGO DA MURTA - Precipitação diária total - Average -1	Average	-1 day
150/02 / REGO DA MURTA - Precipitação diária total	Average	0 day
150/02 / AGR-01 - Nível hidrométrico Instantâneo	Percentil_95	0 day

**Dependent Variables:**

Nome	Agregado	Desfasamento
Dependente - 150/02 / AGR-01 - Nível hidrométrico Instantâneo	Percentil_95	0 day

Powered by SHIRI Data Mining



www.ydreams.com

# Online Data Mining Service (Server's perspective)

---

Meta-Information

Upload

Download

Web Services

---



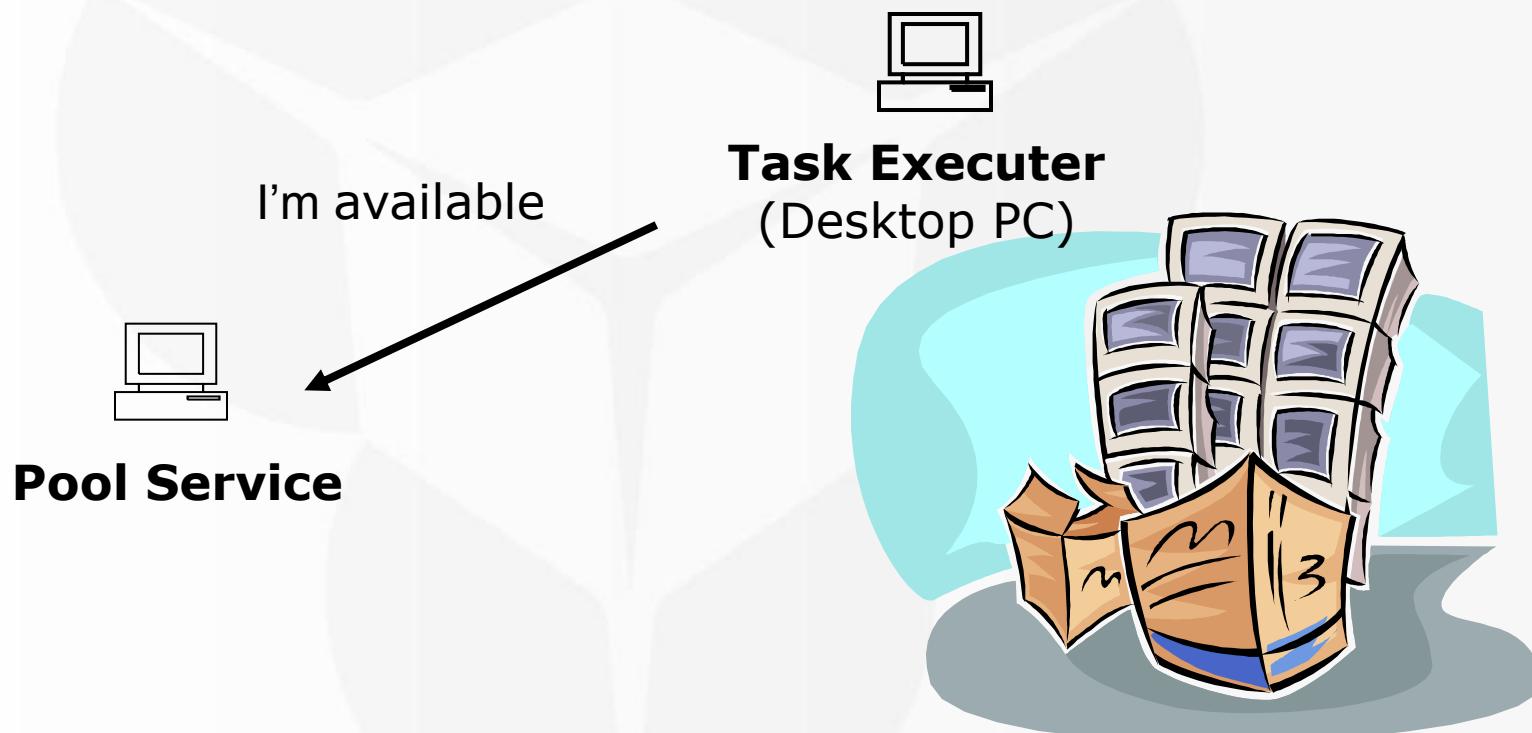
# Online Data Mining Service (Server's perspective)





www.ydreams.com

# Distributed task execution



Meta-Information

Upload

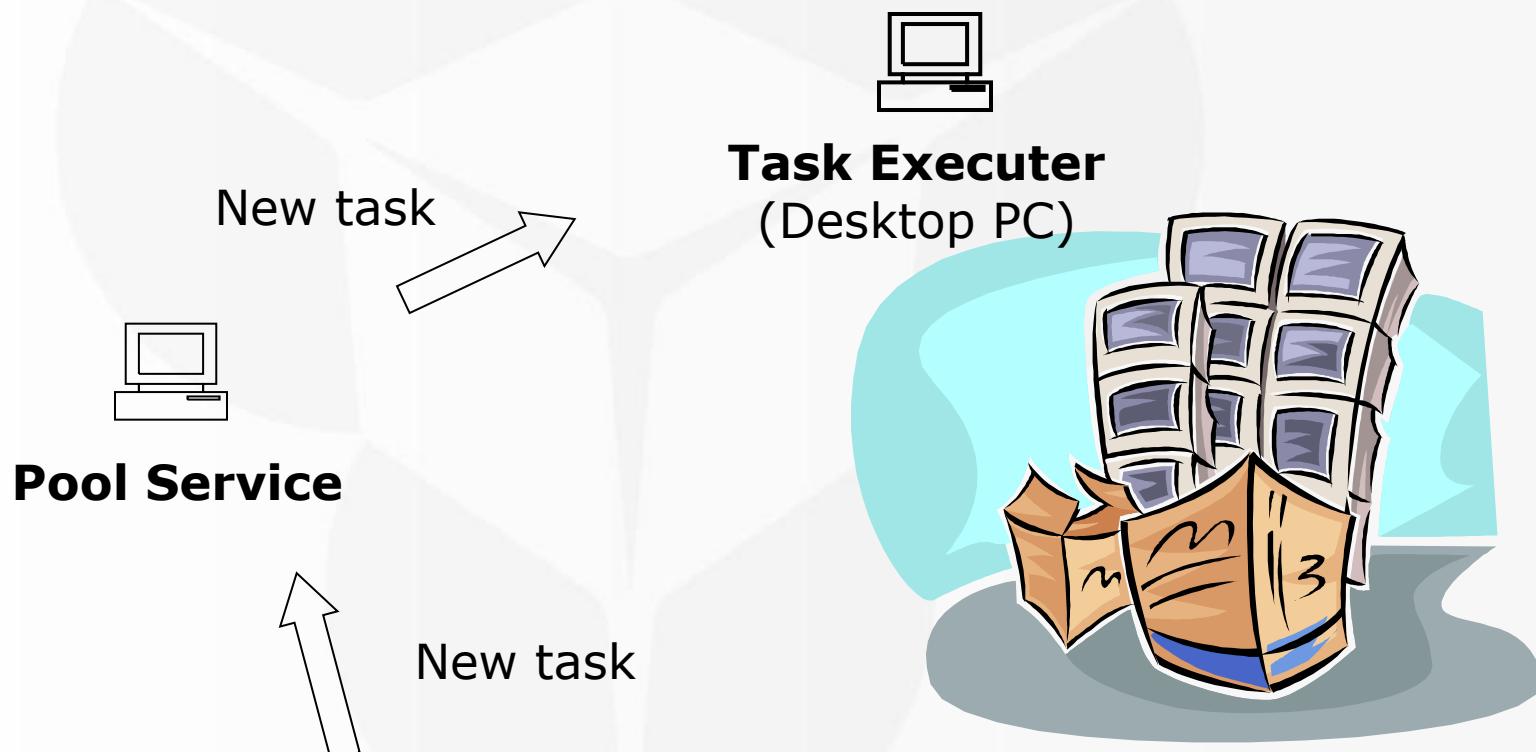
Download

Web Services



www.ydreams.com

# Distributed task execution



Meta-Information

Upload

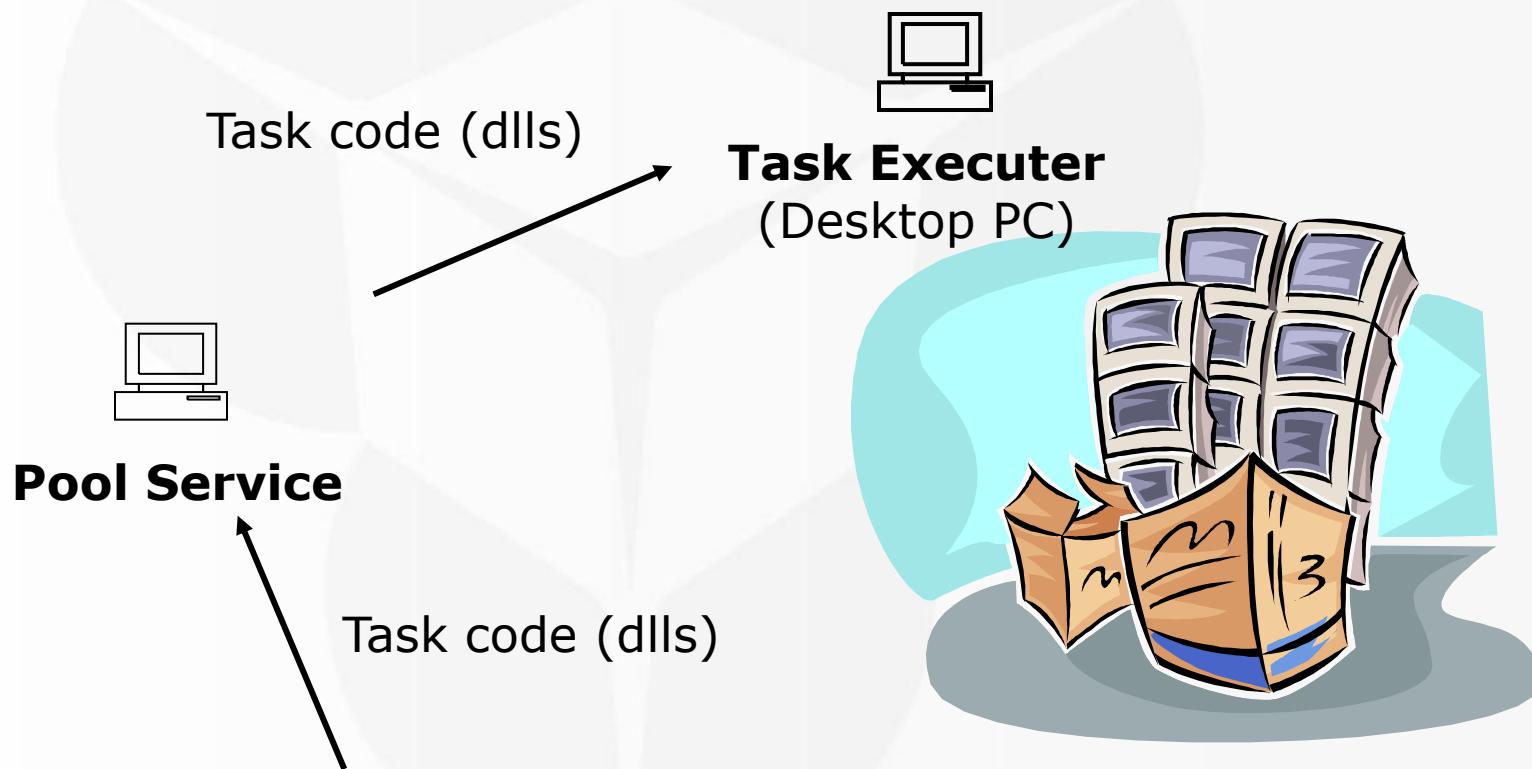
Download

Web Services



www.ydreams.com

# Distributed task execution



Meta-Information

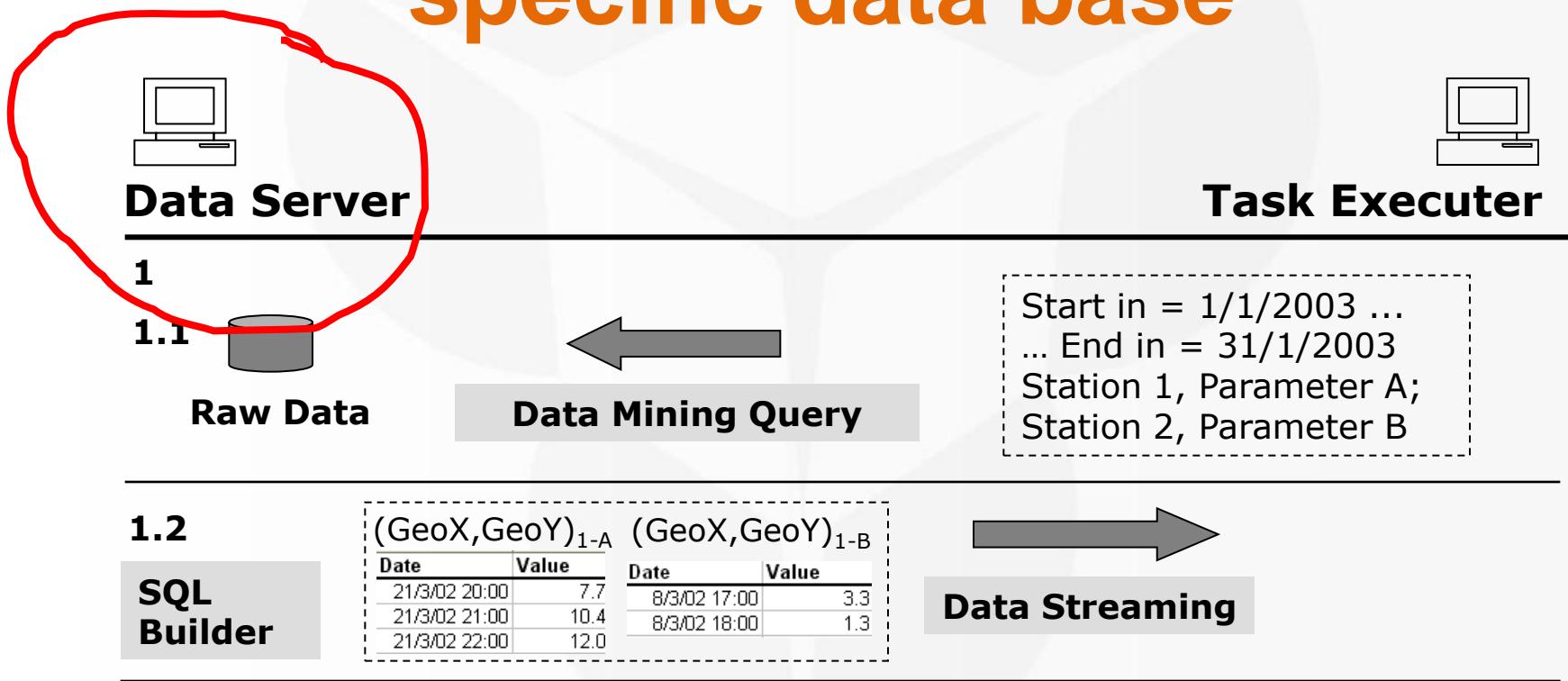
Upload

Download

Web Services



# Customization : targeting a specific data base

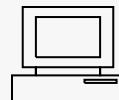




# Customization : targeting a specific data base



**Data Server**



**Task Executer**

**2**

**Data Preparation**

Date	Station 1, Parameter A	Station 2, Parameter B
Percentile 95 per 24 Hours		Percentile 95 per 24 Hours
1/1/2003	1.1	4.3
2/1/2003	2.3	3.1

**3**

**ANN Training**

Station 2, Parameter B
Percentile 95 per 24 Hours
4.3
3.1

$$= f( \frac{\text{Station 1, Parameter A}}{\text{Percentile 95 per 24 Hours}} )$$

Station 1, Parameter A
Percentile 95 per 24 Hours
1.1
2.3



# Spatial Sound Data Mining, the project

Automatic Air Quality Index forecast 1 day ahead of 4 air pollutants (Nitrogen dioxide, ozone, sulphur dioxide, inhalable particles)

The prediction is made available to end users through mobile phones. By using **Location Based Services (LBS)**, the user receives the forecast for the closest monitoring station.



# **Spatial Sound Data Mining, the project**

Applied to the **QualAr** database of the **Instituto do Ambiente**

Kick off of the Online Data Mining Services and a proof of concept



# Spatial Sound Data Mining

QualAr Meta-  
Information Service

Auxiliary  
services

Sound  Maps



Data Mining  
Service

Automatic Services  
(prediction of air quality  
index)

Client Applications



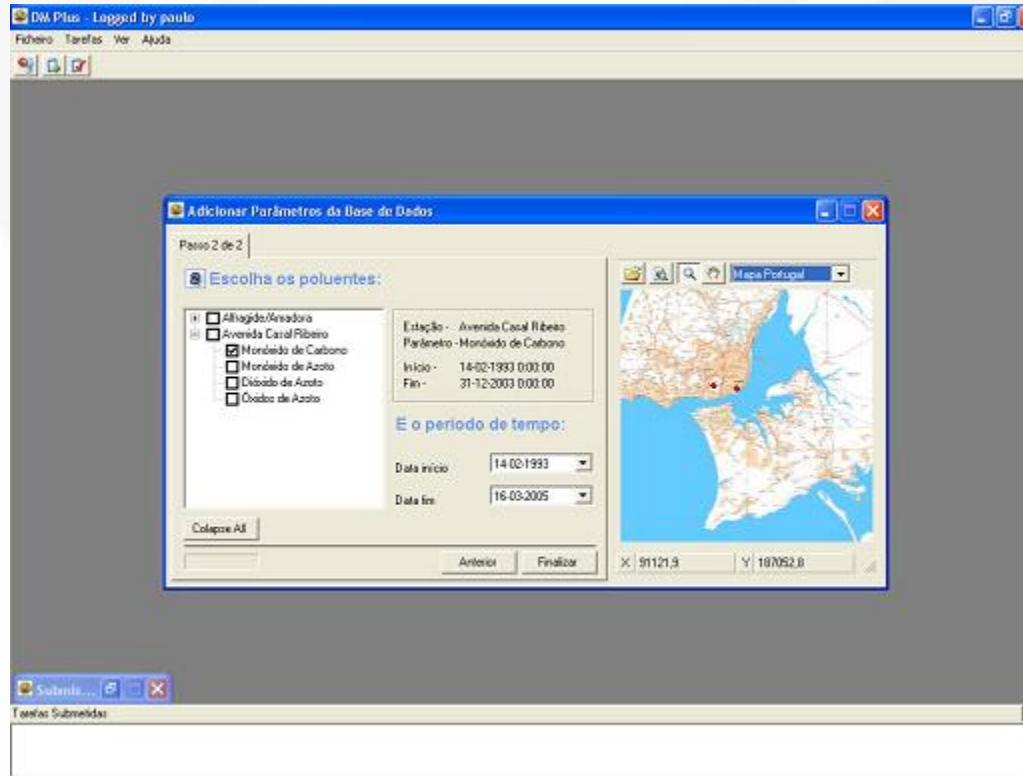
DM Plus QualAr



Automatic Forecast of  
the Air Quality Index



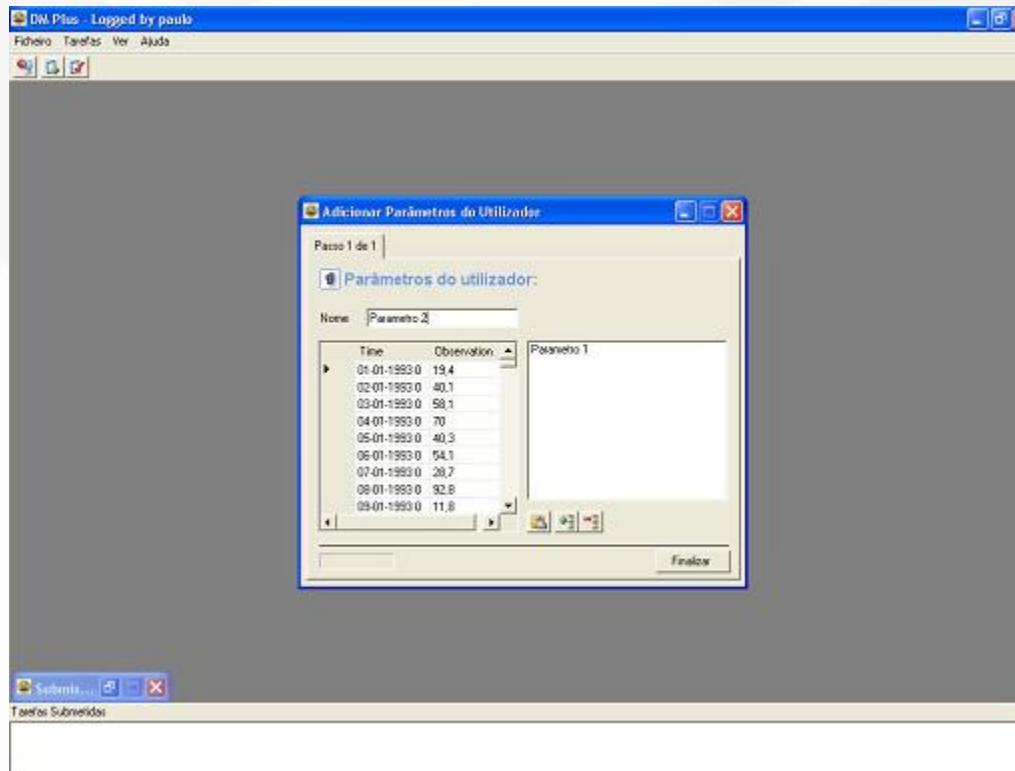
# Spatial Sound Data Mining DM Plus QualAr





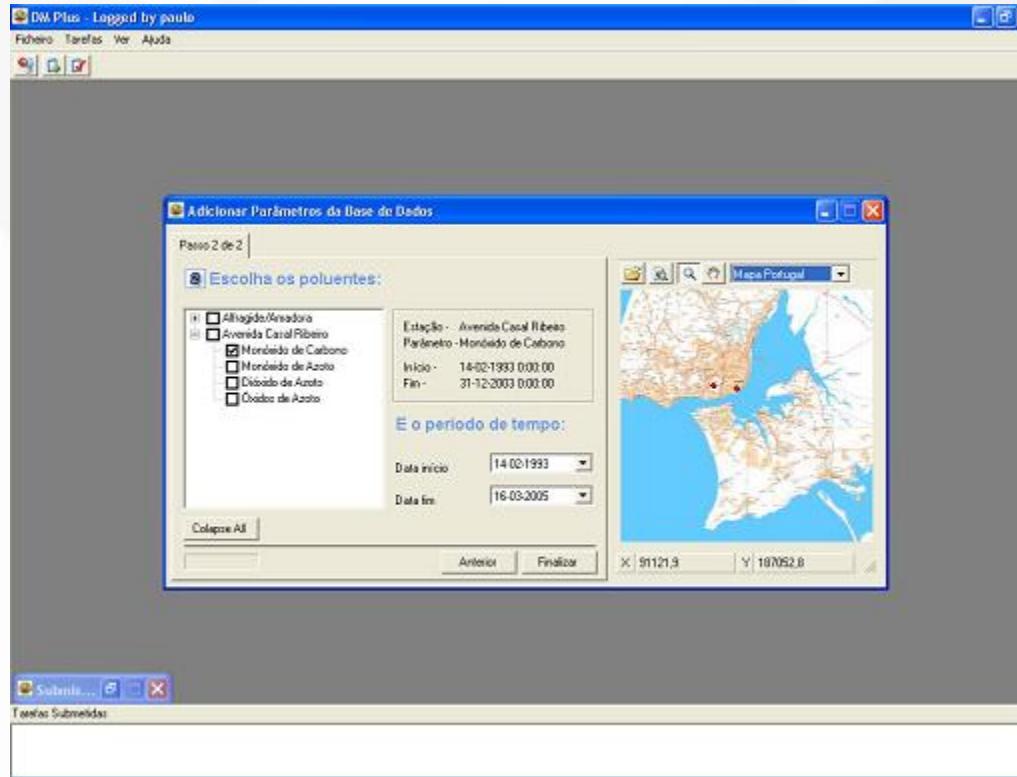
www.ydreams.com

# Spatial Sound Data Mining DM Plus QualAr





# Spatial Sound Data Mining DM Plus QualAr

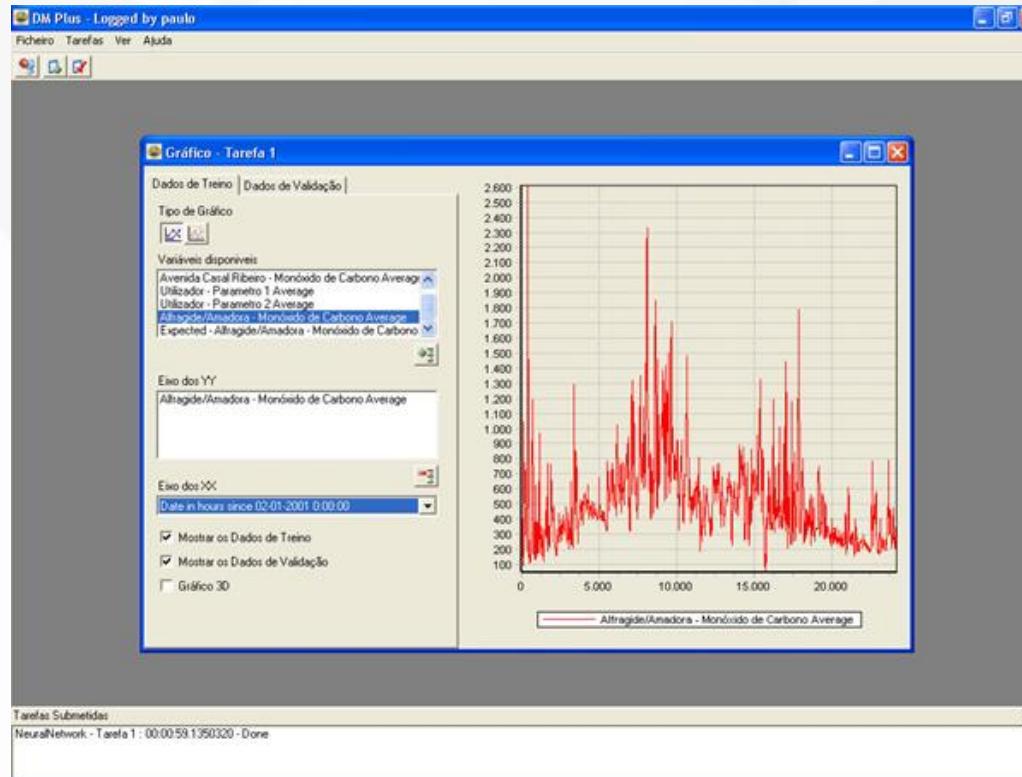




www.ydreams.com

# Spatial Sound Data Mining

## DM Plus QualAr





# Spatial Sound Data Mining

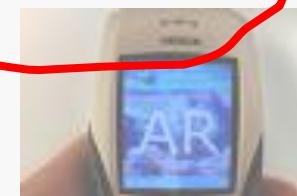
After defining the most relevant  
Input parameters

**Auxiliary  
services**



**Data Mining  
Service**

**Automatic Services  
(prediction of air quality  
index)**



Forecast Service uses automatic trial and error procedures to retrain the networks with new measured data.



# Spatial Sound Data Mining

## Index representation

Limited graphical capabilities and processing constraints of the targeted platform (wireless mobile phone)

Sonification: use of non-speech audio to convey information



# Spatial Sound Data Mining Index representation

Audio used to increase perception of the information:

- the air quality index and each pollutant are associated to a sound.
- *a melody is composed according to a composition rule that specifies the order and duration of each sound.*

**Design challenge:** create a sonification that conveys a relevant subset of the information presented in an acceptable time interval.



# Spatial Sound Data Mining Index representation

Server composed by two main units: Web Server Unit and Synthesis Unit

Client connects (java application) to the air quality service provider and requests information regarding the air quality status.

User listens to the sonification of both the index and its parameters



# Marquis

**European Air Quality database** (6 Countries) and information services for the general public

The online data mining services will be used to interpolate empty records caused by anomalies in the monitoring stations



# Conclusions

The online data mining services applied to environmental monitoring networks have shown to be a quite useful tool to associate with dynamic databases

The online data mining services, like in SNIRH Data Mining, is a new concept of data exploration

The online services to spatial analysis will happen in a near future



# Contacts

[manuel.costa@ydreams.com](mailto:manuel.costa@ydreams.com)

**SNIRH Data Mining**

[marco@inag.pt](mailto:marco@inag.pt)

**Spatial Sound Data Mining Project**

[lmr@porto.ucp.pt](mailto:lmr@porto.ucp.pt), [afonseca@igeo.pt](mailto:afonseca@igeo.pt)