Applying Data Mining to Determine Relationships between Relief and Precipitations: A Case of Study In Northern Chile

Marco Seron[1], Pamela Soto[1], Claudio Meneses[1],
Rodrigo Riquelme[2], and Jorge Quezada[2]
Universidad Católica del Norte, Antofagasta, Chile

[1] Department of Systems & Computer Engineering
{maseron, pamelasoto, cmeneses}@ucn.cl
[2] Department of Geologic Sciences
{rriquelme, jquezad}@ucn.cl

Abstract. In this work we are interested in identifying aspects of the relief that can be associated to tectonic or climatic factors. Applying Data Mining techniques on a group of attributes generated from the transformation of longitudinal profiles of a topography using Wavelet Transform and Matching Pursuit and the historical data of the average precipitations of the last 30 years. The study considers the region among the latitudes 18°S and 27°S in the coastal zone of northern Chile. The objective this study is to find any interesting relationship among latitudinal variations of the characteristics of the relief of the Mountain range of the Coast and the latitudinal climatic variations in northern Chile. The results obtained and an analysis of them indicates no evidence of lineal correlation among relief and climatic variations on this zone, although association rules show certain relationship between relief and precipitations, but with low confidence.

1 Introduction

The role of climate a tectonic on the landscape evolution is widely documented in geomorphology. Tectonic together with climate are the most important factors determining the rate at which a landscape is eroded. In this context, several investigations analyze the landscape topography in order to determinate the relative impacts of these factors in the landscape configuration (e.g. [20], [6], [17]).

Nowadays, availability of Digital Elevation Model (DEM) allows us to explore and to extract some quantitative characteristics of the topography, which can be interpreted in terms of tectonic and climatic factors operating on the landscape [3]. However, the establishment of the impact of climate (or/and tectonic) on the landscape frequently remain uncertain because no formal rules relating landscape characteristics and climate (or tectonic) are used.

In this paper we propose and apply an objective and effective methodology concerning to the discrimination of the relative impact of the climate and tectonic on the landscape configuration. We choice the Coastal Cordillera of the Atacama Desert

© M. Moreno-Ibarra, S. Levachkine (Eds.)
Semantics in Geoinformatics Applications
Research in Computing Science 25, 2006, pp. 87-108

(northern Chile), a mountain range extending more than 1000 km between the latitudes 18°S and 28°S, in the hyper-arid northern Chile. Three factors makes of the Costal Cordillera an excellent place to carry-out our objectives: (1) its topography exposes accurate NS oriented segment which can be, a priori, correlated to the segment of the Atacama Fault System (AFS), a discontinuous set of faults striking parallel to the Coastal Cordillera [22]. (2) It is effected by overall uplift tectonic conditions during the Quaternary. (3) Hyper-arid conditions prevail at least from late Pliocene [13] and, a perceptible NS mean annual rainfall gradient define the present-day climatic conditions of Coastal Cordillera [8].

We address to individualize which of the landscape characteristics of the Coastal Cordillera are related to tectonics factors and which of them can be associated to climatic conditions. To analyze the Coastal Cordillera landscape we consider NS oriented topographic profiles defining mathematical signals, where amplitude corresponds to topographic altitude varying again latitude (from north to south). In this signal, low frequency variations represents variations in large-scale landscape characteristics (e.g. segments of the landscape with different mean altitude). Whereas, high frequency variations represent latitudinal variation of incision, i.e. north-south successions of valleys and interfluves. We suspects that low frequency variations can be related to differential latitudinal landscape tectonic control associated to the different segment of the AFS. Whereas, high frequency variations evidence a greater capacity to incise the landscape that responds to the NS mean annual rainfall gradient or latitudinal variations of the uplift rates: a latitudinal increase of mean annual rainfall or uplift rate inducing an increase of amplitude and frequency of incision.

In order to test our hypotheses, we analyze the topographic signals by considering data transformations performed over a derived group of attributes. These transformations are carried out using the Wavelet Transform (WT) and Matching Pursuit (MP) techniques.

With the topography one can obtain information on the position of an element with regard to the longitude, latitude, and its elevations, among others. The way in that the precipitation erosions a landscape responds to a mathematical model [17]. Based on the mentioned above, we can think that observing the relief, we may determine the magnitude of mean annual rainfall that affected it.

From the Data Mining perspective, one of opposing problems when working with satellite images, is to find some transformation for these images. This transformation should be the more appropriate to extract the information associated to the problem.

For this problem, mainly two sources of information are available: satellite images and data of average coastal precipitations of the last 30 years in northern Chile. With this information, we want to explore the existence of significant relationships between the relief and the precipitations in the region limited between the latitudes 18,5°S (Arica) – 27°S (Caldera), and between the Coast Mountain and the beginning of the Intermediate Depression.

In addition to the main goal of identifying interesting relationships among relief and climate or tectonic factors, we consider the following specific goals:

– To confirm or to discard the hypothesis that the distribution of those systems of drainage of the Coast Mountain are related with the variations of latitudinal precipitations.

- To find any interesting correlations between topography and precipitations in the zone limited from Arica to Caldera and from Coast Mountain to Intermediate Depression of northern Chile.
- To carry out an spectral analysis of the topography, in order to verify whether the topography determines information on precipitations.

2 Methodology

In order to replicate this study to other zones, we formulate a guide (sequence of phases and tasks), based on the methodology propose to carry out a Data Mining project [10]. Our guide (see Fig. 1) is composed of the following phases:

- **Data Preparation:** In this phase, we carry out all necessary data preparation tasks over the two sources of data: satellite topography data and precipitations records corresponding to the target zone. In this phase, the followings tasks should be performed:
  - *Obtaining Longitudinal Profiles:* From the topography data, longitudinal profiles are traced along the Coast Mountain. These profiles must have a constant distance among them, and also the first profile must have a constant distance to the coastal border. This operation is performed by using a GIS tool, storing the results in a data file.
  - *Error Correction:* After the profiles are traced, they may pass on topography defects. Therefore, a process of exploration and correction of these defects should be performed if any defect is detected.
  - *Identification of Structural and Lithological Elements:* It consists of the identification of structural and lithological elements of diverse sizes, belonging to the region under study, with the purpose of establishing their relationships with the results to be obtained.
  - *Interpolation of Precipitations:* Starting with the data points for which precipitations are already known, we should interpolate any required intermediate point between two known points, along the zone under study. This with the purpose of having continuous climatic data that can be subsequently correlated with topography data. Any suitable interpolation method should be considered (e.g., lineal interpolation, Lagrange interpolation).

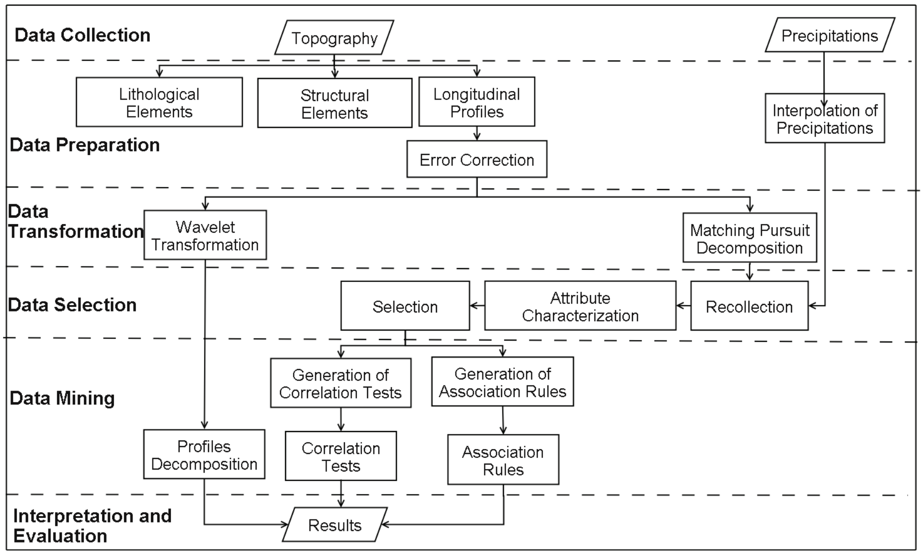

Fig. 1. A schematic overview of the methodology

– **Data Transformation:** Once the profiles are generated, functional transformations are applied in order to decompose them. These transformations map profile data into signal data. In our case, we use the following two transformations:
  - *Wavelet Transform:* The purpose is to obtain a visual decomposition of the profiles.
  - *Matching Pursuit* Decomposition: The purpose is to obtain numeric values of the frequency components represented by the longitudinal profiles.

– **Data Selection:** Once the data have been transformed, it is necessary to identify which one will be finally used in the following phases of the process. Activities that are carried out in this phase are:
  - *Recollection:* The data generated in the transformation and interpolation tasks must be integrated, with the purpose of having a set of possible attributes and records.
  - *Attribute Characterization:* All attributes should be characterized in terms of basic properties such as distribution of values, data quality, statistical information, etc.
  - *Selection:* Once the data are characterized, the relevancy of each attribute should be estimated, using either expert assessment or algorithmic approaches (e.g., filter and wrapper attribute selection methods), or both. In any case, attribute relevancy may depend on the region under study, given that some attributes may be more o less important.

– **Data Mining:** Based on the prepared, transformed, and selected data, data mining techniques are applied in order to generate several candidate models.

In our case, we applied two suitable techniques for the problem of finding relationships in the data:

- *Generation of Correlation Tests:* Statistical tools are applied in order to quantify possible correlations between precipitation data and profile decomposition data, this with the purpose of identifying any dependency between latitudinal variation of the precipitations and the relief of the zone under study.
- *Generation of Association Rules:* Data mining tools are applied in order to generate rules that show significant associations among attribute values. This algorithmic approach may discover no lineal dependency between latitudinal variation of the precipitations and the relief.

– **Evaluation and Interpretation of Results:** The final phase is to evaluate the results (e.g., lineal correlation models, association rules, Wavelet Transform profiles) both individually and crossing the information among them. In addition, it is very important that the results be interpreted by the domain expert and the user in order to have a qualitative estimation whether the results are or not interesting from the expert point of view.

3 An Overview of Applied Techniques

In this section, we overview the main techniques used during the development of the case of study, with the purpose to make the paper as self-content as possible. We describe Matching Pursuit and Wavelet Transform techniques used to decompose profiles (signals), as well as lineal interpolation, lineal correlation, and association rules.

3.1 Matching Pursuit

Matching Pursuit process consists in finding a function f as the pondered sum of elements obtained from a Gabor dictionary [18]. In others words, f is represented as:

$$f = \sum_{\gamma \in \Gamma} \alpha_\gamma g_\gamma \tag{1}$$

This process is used to carry out a decomposition of one signal in different components of frequency that added recompose the original signal. The comparison of the components of frequencies in the Matching Pursuit process is carried out using the Short Time Fourier Transform (STFT) [12], also known as Gabor Transform (GT). STFT cuts the signal in small pieces, which are known as windows. The form of the resulting components of decomposition is strongly bound to the selected window type, which is associated to the fact that the reconstructed signal can have bigger or smaller error.

To decompose this signal, we used the Guimauve software tool [2], which in turn uses the technique of Matching Pursuit to generate the decomposition of the signal into a group of signals located in different time-frequency ranges. Guimauve is based on a group of software libraries, whose name is LastWave [1].

The Guimauve software tool allows to decompose the signal in the following elements:

- **Atoms,** which indicate in how many signals the decomposition of the original signal was carried out. For example, if the quantity of atoms is 200, this indicates that the signal was decomposed in 200 signals.
- **Coefficient,** which shows the coefficient of decomposition of the signal using STFT.
- **Octave,** which is an interval of frequencies in a power of two.
- **Time,** which indicates in that time a certain frequency is found.
- **Frequency,** which indicates the frequency of the decomposed signal.

3.2 Wavelet Transform

The Wavelet Transform [16] carries out a filtrate of the original signal, using filters that allow us to pass a certain range of frequencies either below a threshold (low filters) or above that threshold (high filters) [2], and that eliminate certain components of frequency of the signal. This procedure is repeated with the signals generated from the previous filtrate. The process is carried out mainly with two approaches: The Continuous Wavelet Transform (CWT) and the Discrete Wavelet Transform (DWT).

- **CWT:** It uses a window that frames the signal inside an interval, concentring the analysis only in a segment of the signal. It tries to express a continuous signal X(t) in the time, by means of an expansion of proportional coefficients to the internal product between the signal and different scaled and transferred versions of a prototype function Ψ(t) (mother wavelet) [9]. The CWT is defined as:

$$CWR\ (a,b) = \frac{1}{\sqrt{a}} \int X(t) \Psi \frac{(t-b)}{to} dt \qquad (2)$$

  where: a controls the width of the function (scale variable) Φ(t) and b gives us the location in the domain of the time (translation variable). The scale variable allows us to dilate or to compress the mother function to control the resolution that is wanted to obtain in the transformation. The translation variable controls the location of the function Ψ(t) in the time space. This allows us to displace the mother function of scale a through the signal. With this one can say that to smaller size of a (small scale), a bigger quantity of translations must be performed, in order to cover the complete signal.
- **DWT:** This transformed part of the base that the scale parameters and translation are discrete, this with the purpose of simplifying the quantity of calculations to carry out. Under this assumption, the following thing is generated:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}} \Psi \frac{(t-b)}{to} to, b \in \Re, a \neq 0 \qquad (3)$$

with $a=2^{-j}$ and $b=k2^{-j}$, it becomes the group of functions:

$$\Psi_{i,k}(t) = 2^{\frac{j}{2}} \Psi(2^{jt} - k)\, j,k \in Z \tag{4}$$

This decomposition generates a set of elements, d12 to d1, from small to big frequencies, and a12 corresponds to the first decomposition (also called base decomposition).
Once we know the definitions of both ways of transforms, it is necessary to define the mother function more appropriate for the problem.

The advantage of Wavelet over STFT for this study is that STFT lacks of properties (orthogonality, conservation of the energy, entropy) necessary to be able to make a correct interpretation of the decomposition. To decompose the signal by means of Wavelets, we used Wavelet Toolbox of Matlab 7.0. On the other hand, Wavelet and Matching Pursuit are complementary techniques for this study. The first one allows us to visualize constituent frequencies of the train of signals that can be quantified through the second one. The visual results obtained with Wavelets are corroborated by the quantitative results obtained with Matching Pursuit.

3.3 Correlations

Correlations are used to examine the degree of similarity of the values of two numeric variables. A standard formula to measure lineal correlation is the correlation coefficient r, which is a numeric real value between -1.0 and 1.0. If r is positive then the variables are correlated positively, i.e., the variables have a similar behavior (both growing or falling at the same time); if r is negative then the variables are correlated negatively, i.e., when a variable grows the other one falls [14]. On the other hand, when r is 0, there is not correlation.

3.4 Association Rules

They express behavior patterns among the data in function of the combined appearance of values of two or more attributes. The main characteristic of these rules is that they try with nominal attributes, to difference the correlations that make it with the numeric ones. In short, these rules express combinations of values of the attributes that happen more frequently. They can be expressed in many ways, although the most common form is of the style: “if the attribute X takes the value d then the attribute Y takes the value b”, which usually is written as: X=d → Y=d. An association rule is usually worked with two measures to know the quality of the rule [14]:

- **Support:** It is defined as the number of instances that the rule predicts correctly.
- **Confidence:** Also denominated precision, it measures the percentage of times that the rule is met when it can be applied.

3.5 Interpolation

– **Lagrange Interpolation:** Its is defined of the following form:

$$p(x) = y_0 l_0(x) + y_1 l_1(x) + ... + y_n l_n(x) = \sum_{k=0}^{n} y_k l_k(x) \tag{5}$$

where $l_0, l_1, ..., l_n$ are polynomials that depend only on the sampled values $x_0, x_1, ..., x_n$, but they do not depend on the values $y_0, y_1, ..., y_n$. The general formula for the polynomial li is:

$$l_i(x) = \prod_{j=0, j \neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{6}$$

Using these polynomials in the equation (5) we obtain the exact form of the Lagrange interpolation polynomial [4].

– **Lineal Interpolation:** It is the adjustment of one straight line between two known points, to determine the value of an unknown point that is among both previous ones. Be F(x) a function for we know their values $f(X_0)$ and $f(X_1)$ in the points X_0 and X_1. The more simple approach to approximate the f(x) values for any x belonging to the interval $[X_0, X_1]$, is to consider the polynomial of first degree [4]:

$$F(x) = f(x_0) + \frac{x_1 - x_0}{x_1 - x_0}(x - x_0) \tag{7}$$

4 Sources of Information

4.1 DEM

A DEM [19] is a representation of the landscape of the form **(x, y, z)**, which correspond to the latitude, longitude, and elevation, respectively. The precision of these data depends on two factors: the horizontal precision and the vertical precision.

– **Horizontal Precision**. It corresponds to the resolution of the satellite image from which the topography was obtained. In the case of the DEM used for this study, it has a resolution of 90 meters. That means that each pixel of the image represents a square of 90 squared meters.

– **Vertical Precision**. It corresponds to the precision calculated from data obtained in landscape of elevations, which are compared with the data of the elevations obtained from the DEM. This operation is carried out using the Medium Squared Error (MSE), with at least 30 points, 20 of which correspond at the interior of the DEM, and 10 to those correspond to the border of the same one.

$$MSE = \sqrt{\frac{\sum (Z_i - Z_t)^2}{n}} \tag{8}$$

From the DEM data, we carry out longitudinal profiles in four places, in order to appreciate the variation of the relief when we advance from the coast toward the intermediate depression. These cuts are shown in Fig. 2A, where the first, second, third and quarter profile are shown from left to right. With the information obtained from the profile, we can build scatterplot graphics among the latitude in a point (axis X) and the elevation of the same one (axis Y), resulting in the dispersion graphics shown in Fig. 2B.

These graphics show the form of the relief through the cuts, and where depressions, plains and elevations are appreciated. These profiles can be considered as signals. This since the elements that compose it can be transformed to homologous elements, but in the domain of the signals, and where the elevation represents the width, the latitude can be homologated with the time, and the variations in the elevations of the relief indicate different frequencies in this signal, to which we will call topography.

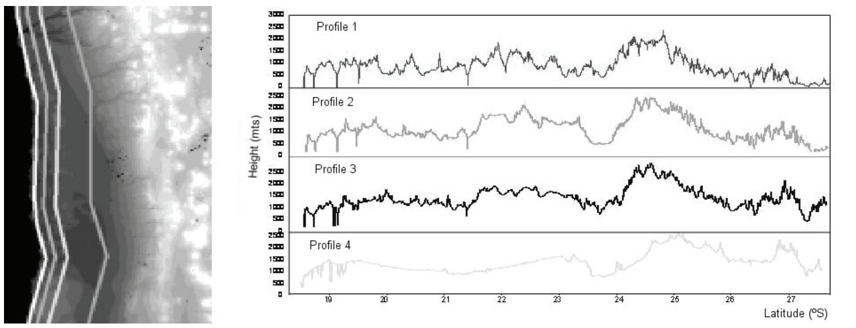

Fig. 2: A) Longitudinal profiles used in the study, B) Scatterplots of the longitudinal profiles.

4.2 Precipitations

The data associated to precipitation correspond to coastal cities located among latitudes 18,5°S and 27°S (see Table 1). These data of precipitations are an average of the last thirty years [8]. Given that the necessary data should be from the latitude 18,5°S until the latitude 27°S, we carry out an operation of interpolation of precipitations in all range of latitudes, with the purpose of being able to use these data to establish correlations and associations.

Over this dataset of precipitations, a lineal interpolation (7) is performed among the diferent data points. In this case, the Lagrange interpolation generates points with smaller values than those in the Table 1. As shown in Fig. 3 (a), the curve generated by the Lagrange interpolation gives negative values between 20°S and 22°S, and some points (e.g., Tocopilla and Chañaral) are quite different from the true ones. On the other hand, as shown in Fig. 3 (b), lineal interpolation does not generate negative values, and it fits the initial values. For this reason, the interpolation of precipitation values is performed by using the lineal method.

Table 1. Precipitations in different cities of the coast of northern Chile.

City	Latitud(°S)	Precipitation (mm)
Arica	18,5	0,5
Iquique	20,8	0,6
Tocopilla	22	2
Antofagasta	23,5	1,7
Taltal	24,4	8
Chañaral	26,2	12
Caldera	27	27
		Mean: 4,13 Standard Deviation: 4,75

5 Data Preparation

Data preparation consists basically in taking topographic and precipitation data with the purpose of generating a consistent table of data and in this way to avoid the inclusion of data that can influence negatively the results obtained.

- **Lithography**. Since this element determines the behavior of the landscape with regard to the precipitations and it influences the form that is modeled.
- **Structural elements**. Since the Atacama Fault crosses the profiles and it affects the formation of the relief in direct form.

Next, we describe the sources of data and the form that these are preparated:

1. **Preparation of Precipitation Data**. We applied lineal interpolation to the precipitation data to approximate unknown values. As we can observe, the lineal interpolation is well adjusted to the group of precipitation data (see Fig. 3 and Table 1).

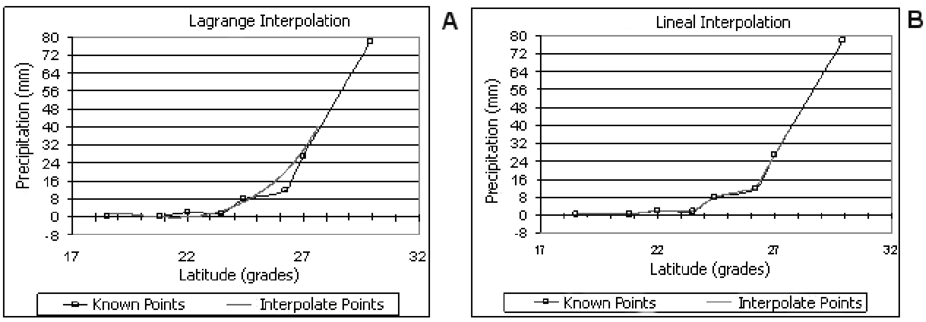

Fig. 3: (A) Scatterplot of the lagrange interpolation of the precipitation data. (B) Scatterplot of the lineal interpolation of the precipitation data.

2. **Determination of lithological elements**. As previously mentioned, the signal “Topography” is made up of different elements that can be reflected in different components of frequencies. Among these elements are rocks of diverse sizes. Also, maps were generated to identify bodies by using convolution filters. Specifically, we used the filter of Means to highlight the bodies of a certain size. The filter carries out a substitution of the value of a pixel for the mean of their k-nearest neighbors, and where the size of the bodies that we want to identify determines the value that k should take. After obtaining the result of the filter application, a subtraction is made among the resultant map and the original one. This operation generates a map that highlights the bodies of certain size. These outstanding bodies are associated to certain type of rocks in the lithography of northern of Chile.
3. **Determination of structural elements**. In the same form as some bodies of different sizes was identified, it is necessary to isolate the components of frequency of the associated signal “Topography” to elements of structural control, specifically the faults, which are discontinuities formed in the rocks for fractures when concentrations of tectonic forces exceed the resistance of the rocks, that generates a relative displacement of the separate parts ([11], [26]).
4. **Determination of the vertical error**. The vertical error consists on the error introduced when calculating the elevations in the generation of a DEM from a satellite sampling. This error depends specifically of the satellite that carries out the mensurations, in our case the GTOPO90. This error takes place when for atmospheric, climatic factors, floor types and other reasons, radio’s waves sent toward the earth are deviated of the course that should have under good conditions, this variation generates the differences of elevations. These vertical errors are estimated ([25], [7], [15], [5]) of the order of ±25 meters.
5. **Correction of longitudinal profiles**. In the generation of the profiles, we detected the presence of sectors in which the elevations were zero, this corresponds to errors in the generation of the base satellite image. This was detected in the exploration of the data of each profile. To overcome this problem, we calculated the mean value between the borders of the hole, and we put this value in the corresponding position of the missing values. Afterwards, the same operation is carried out between the border and the inserted value. In this way, we fill in missing values of the generation of the profile.

6 Data Transformation

The decomposition of the signal “Topography” seeks to identify which of the obtained components of frequency is related with the latitudinal precipitations in the region mentioned before. The signal decomposition process was carried out by means of the use of Matching Pursuit and Wavelets methods, with the purpose of comparing the results of both methods. We used Matching Pursuit for the evaluation of the association rules and correlations, by means of the Guimauve tool. When STFT is used it should be indicated which is the type of sampling window necessary to generate the dictionaries of signals.

We carried out a group of decompositions of the signal corresponding to profiles 1, 3 and 4, in way of seeing which are the errors generated to the moment of reconstructing this signal. In Fig. 4 we can see the reconstruction error for the first profile. The objective of this test is to determine which is the more appropiate window for the decomposition of the signals

The behavior of the error shown in Fig. 4 is similar in all the analyzed profiles. Also, we can observe that the Blackman window type is the one with worse results, and that the Hanning and Spline 1 window types show the best results when trying to reconstruct the original signal.

Finally we use the Spline0 type, that corresponds to the family of the B-Spline and is known as rectangular ([24], [23]). The rectangular form of this window allows us to appreciate clearly signal components along the time, without deformations, which facilitates the task of interpretation of each component.

We used Haar Wavelets to identify in to graphic way the “blocks” of frequencies. This Wavelet mother facilitates the visual comparison of the decomposition with those “blocks” of frequencies of the topography and it also allows to compare it with the decomposition accomplished by means of rectangular Spline in Matching Pursuit. The tool used to generate this decomposition was Wavelet Toolbox of Matlab 7.0.

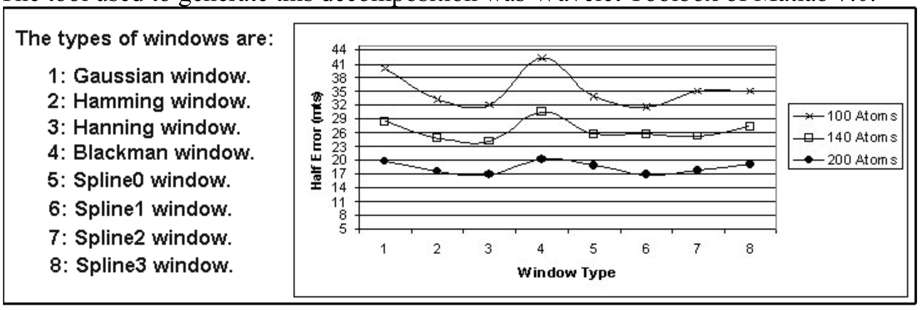

Fig. 4: Error for the different windows.

7 Data Selection

7.1 Compilation of Attributes

As a result of all the mentioned above, we obtain the following attributes:

- **Longitude (L).** It corresponds to the longitudinal location of an area, inside the study area.
- **Initial Latitude (LI).** It corresponds to the initial latitudinal location of a component of frequency.

- **Final Latitude (LF).** It corresponds to the final latitudinal location of a component of frequency.
- **Latitudinal Distance (LD).** It corresponds to the difference between Initial and Final Latitude, and it is measured in degrees.
- **Frequency (F).** It corresponds to the value of a component of the signal.
- **Precipitation (P).** It corresponds to the quantity of precipitations in an area of the study region. They are measured in millimeters of water fall.
- **Elevation (E).** It corresponds to the values of the elevations in the study region, measured in meters.

7.2 Characterization of Attributes

The candidate attributes are characterized of the following form (see Fig. 5):

A

Name	Type	Units	Possible Values
Longitude	Numeric	Degrees	Min.: 69,7767 Max.: 70,8482
Initial Latitude	Numeric	Degrees	Min.: 18,3351 Max.: 27,5221
Final Latitude	Numeric	Degrees	Min.: 18,4227 Max.: 27,5278
Latitudinal Distance	Numeric	Degrees	Min.: 0,0024 Max.: 5,8619
Frequency	Numeric	No unit	Min.: 0 Max.: 0,25
Precipitation	Numeric	Millimeters	Min.: 0,503868 Max.: 36,23628
Elevation	Numeric	Meters	Min.: 40 Max.: 2650

B

Attribute	Mean	Standard Deviation
Latitude	23,1889508	2,714656427
Precipitation	8,402134897	10,74030827
Frequency	0,011599665	0,011238048
Elevation	959,5321571	482,6221739

Fig. 5: (A) Maximum and Minimum values of Attributes, (B) Descriptive Values of Attributes.

In the exploration of the data of the profile corresponding to the coast (first profile), it is possible to determine the distribution of the values for the following attributes: latitude, precipitation, frequency, elevation (see Fig. 5 (B)). Standing out the fact that the elevation has a near behavior to a normal distribution in the complete range of the latitudes. On the other hand, the precipitation and the frequency have a exponentially falling distribution. The maximum and minimum values of these attributes can be seen in Fig. 5 (A).

7.3 Selection of Attributes

The criterion used to select attributes was to discard attributes with direct relationship to the geographic position. This pursues to avoid to introduce bias in the generation of models.

Given this, four attributes were selected: Frequency (F), Precipitation (P), Elevation (E), and Latitudinal Distance (LD).

8 Data Mining

8.1 Association Rules

For all the models, we considered 600 instances corresponding to the union of the decomposition of the three nearest to the coast profiles. Each profile was decomposed in 200 atoms. Table 2) summarizes association rules generated by using the WEKA [27] software tool. This rule indicates the presence of some associations among the attributes that could show some relationship between relief and climate. In this model interesting rules are generated that associate the Latitudinal Distance, the Frequencies and the Elevations in the first rules (we show only some rules that indicate the presence of relationships that we are looking for).

Table 2: Association rules found to relate attribute values.

Rule number	Antecedent	Consequent	Confidence
17.	F=middle-high LD=low E=high **17**	P=middle **15**	acc:(0.84262)
18.	F=middle E=middle-under **23**	LD=low **20**	acc:(0.83892)
19.	F=middle-high E=high **23**	P=middle **20**	acc:(0.83892)
20.	E=middle-under **123**	LD=low **101**	acc:(0.80862)
21.	F=middle **120**	LD=low **96**	acc:(0.78732)
22.	E=middle-high F=low **22**	P=low **18**	acc:(0.78646)
23.	E=middle-high LD=middle **16**	P=low **13**	acc:(0.7717)
24.	F=middle-under E=low **16**	LD=low P=high **13**	acc:(0.7717)
25.	E=middle-high **120**	LD=low **92**	acc:(0.75974)
26.	F=low E=high **23**	P=middle **18**	acc:(0.75466)
27.	P=middle **203**	LD=low **151**	acc:(0.74333)

8.2 Correlations among Profiles

To quantify possible lineal relationships among the attributes mentioned above, we computed the Pearson correlation coefficient for each one of the three first profiles separately, obtaining the following results:

- In Tables 3A, 3B, and 3C, we can observe that the lineal correlation between frequency and precipitation is growing among the profiles, i.e., the first profile has smaller correlation than the second and the third ones, in the same way that the second profile has smaller correlation than the third profile. It was expected that the results were inverse, i.e., that the correlation among precipitations and frequency were bigger in the first profile than the others, and it was falling as the profiles go into the mountain range, from the coast to the intermediate depression. This because the available precipitation data are only associated to the coast (first profile).

- In Tables 3A, 3B, and 3C, we can also observe the existence of a medium negative correlation among elevation and precipitation. This correlation increases as the profiles go into the mountain range of the coast.

Table 3. Correlations of Profiles

First Profile (A)

LD	1			
F	-0,197	1		
P	-0,191	0,039	1	
E	0,208	-0,07	-0,396	1
	LD	**F**	**P**	**E**

Second Profile (B)

LD	1			
F	-0,203	1		
P	-0,197	0,07	1	
E	0,12	0,063	-0,386	1
	LD	**F**	**P**	**E**

Third Profile (C)

LD	1			
F	-0,2	1		
P	-0,275	0,337	1	
E	0,23	-0,28	-0,487	1
	LD	**F**	**P**	**E**

8.3 Maps to confirm the presence of tectonic and lithologic elements in low and medium range frequencies

It had been determined previously that the components of low frequency were associated to structural factors (basic tectonic configuration), as illustrated with an example in Fig. 6A and Fig. 6B. Then it was necessary to corroborate that the medium range frequencies are associated to types of rocks and geologic bodies. As shown in Fig. 6C, medium range frequencies can associate to bodies and types of rocks according to their size. To obtain these results we applied the technique of subtraction of maps of convolution of medium range elevations.

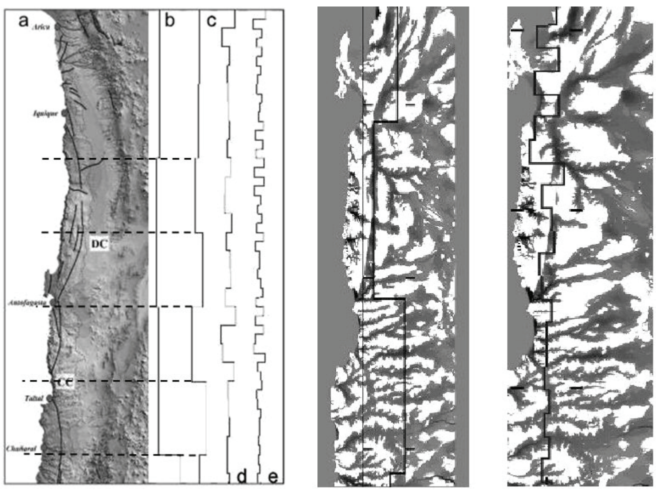

Fig. 6: A) General scheme [a) Relief, b) decomposition a1 (it bases) three blocks are appreciated along the whole region, c) decomposition d12 shows sections of the Fault of Atacama, the decompositions d10 (c) and d9 (d) you show elements lithologic of different magnitudes], B) Convolution Map, matching frequencies with faults, C)Convolution Map, matching frequencies with lithologic elements.

9 Interpretation and Evaluation

Based on the results obtained, we can make the following observations:

1. ***The existence of a relationship between relief and climate of latitudinal form in northern of Chile can not be determined in a precise way.*** **(R.1)**. This because the proposed models have results that conduct to contradictory conclusions.

   (a) On one hand, a contradiction takes place among the correlations of the profiles considering frequencies and precipitations (see Tables 3A, 3B, and 3C). It is obtained that the third profile (the farest one from the coast) has a correlation coefficient of *0,337*, which is bigger than the correlation obtained for the first profile (0,039) (the nearest one to the coast). Considering that the precipitation available data correspond only to the coast of northern of Chile, i.e., to the first profile, the corresponding correlation coefficient does not allow us to find evidence of the existence of lineal correlation among frequency and precipitation.

   (b) The association rules obtained indicate the existence of a relationship among a range of frequencies corresponding to the topography and the precipitations, but these rules have low support and confidence values (See Table 4).

Table 4. Association rules that relate Frequency and Precipitation values.

Antecedent	Consequent	Confidence
F=high 40	P=medium 26	acc:(0,635)
F=medium-high 44	P=medium 20	acc:(0,445)
F=medium 40	P=medium 17	acc:(0,416)
F=low 36	P=low 15	acc:(0,406)

   (c) Given the results R.2 and R.3 that are detailed to continuation, we eliminated of the space of search components of frequency associated to structural or lithological factors. The component range of frequency (high frequencies) remaining could contain information associated to the latitudinal variation of the precipitation. But the amplitude of these components of frequency are in the margin of the vertical error associated to the source of the topography. For this reason, we cannot discriminate which values can correspond to errors or to effect of the precipitations. This because doesn't exist yet a theoretical form to discriminate with confidence elevations in the error range.

2. ***We determine a relationship among components of frequency of the topography signal and elements of the basic tectonic configuration.*** **(R.2)**

   Three components of the topography signal are determined (low frequencies) associated to elements of the basic tectonic configuration. This through the decomposition via Wavelet and later on corroborated by STFT. An example of this can be appreciate in Fig.s 6B and 6C. We determined that the Wavelets decompositions corresponding to the blocks a12, d12 and d11 are related to structural factors. We associated those blocks with the latitudes corresponding to each section, which correspondence of these elements was corroborated by geology experts by means of visual inspection. The verified structural elements correspond to the Atacama Fault in their three main segments: Salar del Carmen, Paposo, and Vicuña [21]. Also, we detected the presence of a block corresponding to inverse faults in the sector of the first region of Chile (Tarapaca).

3. ***We determine a relationship among components of frequency of the topography signal and lithology, together with geologic bodies in the study area.*** **(R.3)**

   A group of components of the topography signal is determined (medium range frequencies) that are associated to lithographs and geologic bodies at different scales. These bodies ranges from the 40 kms. to the 3.3 kms. of longitude approximately. They were detected using the Wavelet decompositions, which corresponds to the d10, d9, d8, d7, and d6 components. This was corroborated by using convolution maps and visual inspection of the correspondence of the components of frequency and the maps (see Fig. 6B and 6C).

4. ***We determined a medium-low negative correlation among the latitudinal variation of precipitations and the elevation of the relief.*** **(R.4)**

   When carrying out the correlation among the variation of the precipitation and the variations of the latitudinal form of the elevations, we obtain a correlation coefficient of -0,396 (see Tables 3A, 3B, and 3C). In spite of being a small value, it indicates a tendency that the geology expert considers interesting and that can be explored in later studies.

10 Discussion

10.1 Revision of the process

The process carried out in this study consisted mainly of two stages:

1. Generation of the data, which in turn is decomposed in the following steps:
   (a) First, we used GIS tools was used in order to generate from a DEM a longitudinal profile that corresponds to the topography signal in the study.

   (b) Later on, this signal is subjected to decomposition using Matching Pursuit (MP) with the Guimauve tool and decomposition via Wavelets with Matlab 7.0 tool.

   (c) With the data generated by the decomposition, we built a data table, in which were determined attributes that characterize the signal, and other additional attributes that characterize the relief and the precipitations.

2. Application of Data Mining techniques:

(a) With data table generated, it was carried out diverse discretizations on the attributes to be able to determine which one distributes the data to be studied in a better form.

(b) We carried out correlation tests among the different attributes to determine if they have some relationship type among them.

(c) We generated association rules using the algorithm PredictiveApriori algorithm implemented in the WEKA software tool, and whose results were analyzed and commented together with the results of the correlations.

This process and work can be improved in several forms. For example, we can improve the quantity and quality of the obtained data, as well as the use of other tools. In order to be able to discriminate the range of frequencies that currently are not totally analyze yet, we identify some changes or improvements that can be implemented in a following phase of this study:

1. The Guimauve tool carries out the decomposition of the signal through MP using a dictionary of STFT. However, this transformation may not be the most appropriate tool to carry out this work, given the asymmetric and acyclic characteristics of the signal. On the other hand, the Matlab software carries out the decomposition of the signal by using Wavelet, but their results are images, and it is not possible to determine quantitatively the frequency of the signal, which is very important for this study. Afterwards, the results obtained with Guimauve were graphically compared to the results obtained with Matlab. In the future, signal decomposition can be made by using software libraries, as alternative to the use of the software tools mentioned above.
2. For the data of the precipitations the available information corresponds to the average of the last 30 years of 8 coastal cities to carry out the interpolation. This restriction may introduce noise in the derived precipitation data, which in turn may affect the precision of the results. Therefore, it is necessary, in order to establish clearly the effects of precipitation over relief and vice versa, to have a bigger quantity of precipitation data, in particular coastal precipitation data and precipitation data associated to the coastal mountain range. This will allow us to carry out comparative studies between coastal mountain range and internal mountain range.
3. It is necessary to have a more precise map of rock types, since we used a simplified map that classifies the rocks for geologic time and not for geologic type. This will become necessary to be able to determine weather or not exist components of frequency associated to the rock type.
4. A parameter that may be tuned corresponds to the quantity of atoms in the decomposition of the signal using Guimauve. This because it may be feasible that a bigger quantity of atoms improve the results. Together with this, data discretization may affect the quality of the results obtained, and therefore it is an issue to be evaluated.

5. The lithological and structural components of the relief should be addressed in a future study, this with the purpose of minimizing the quantity of variables that may influence in the problem.

10.2 Expert Evaluation

The results obtained were presented to the expert, who made the following conclusions:

1. Apparently, the precipitations don't have relationship with the relief in the region under study (northern of Chile).
2. Given this, we can affirm that the climate has not varied significantly in the last 20 million of years. This conclusion is reinforced with studies of superficial datation in the area of the study. These studies date surfaces in northern Chile to that antiquity. This is related because if exists a bigger quantity of precipitations in the past these surfaces could not have been conserved. The sentence mentioned above is explained since if has existed in some moment inside the 20 million years a bigger quantity of precipitation, the landscape should not have been conserved. This is based on that to bigger quantity of precipitations, a bigger evolution of the landscape takes place.
3. This implies that the variations of the relief in the north area of Chile, you it is dominated by the basic tectonic configurations of the place and I didn't seize for the precipitations.
4. Is considered successful the carried out work for that although, it was not gotten to demonstrate the relationship between the precipitations and the relief. Data were gotten excellent about the factors that determine the relief. As well as it was proven that one can to work with ghastly analysis for analysis of topographies.

In summary, this work presents an application of the data mining process to the geologic domain, where data require transformation before applying them to the appropiate techniques. In this case, techniques require to transform the data to the frequency domain, where decomposition can be applied in order to identify components of the original signal. As shown, traditional statistical techniques (e.g., lineal correlation) and machine learning approaches (e.g., association rules) can be used to correlate and associate attributes values. In the case of association rules, attribute discretization must be applied before running the algorithm.

The user evaluation and interpretation of the results indicate that the relief in the zone under study (northern of Chile) is dominated mainly by tectonic variations and not for climatic variations.

11 Conclusions

The quantitative results obtained of the application of Matching Pursuit are not correlated for any constituent frequency with the latitudinal variations of the precipitations. However, other methods of establishing dependencies among data, specifically associations rules, show correlations among precipitation data and topographic latitudinal profiles. This is also corroborated with the visual results obtained of the application of the Wavelet Transform. Indeed, the decomposition with WT indicates us that the lithologic and tectonic factors are of first order in the control of the topography of the Coast Mountain in northern Chile. Thus, the constituent lower frequencies are controlled by the main segments of the AFS (Fig. 6 .A.a and 6 .A.b), while, those higher constituent frequencies are related with lithologic elements of diverse sizes (Fig. 6 .A.c and 6 .A.d). The highest frequencies can be associated to lithologic bodies until their amplitude (local relief) arrives to the margin of the vertical error. Therefore, the hypothetical existence of climatic control in the configuration of the relief of the Coast Mountain range, should be verified for morphologic features smaller than those that are possible to observe with GTOPO 90.

The application of this methodology allows us the study of factors involved in the evolution of the relief in a more analytic way together with outlining an approach based on the treatment of signals in the analysis of a relief. The combined application of the techniques of Transformed Wavelet and Matching Pursuit for the analysis of the relief allows us the identification of tectonic and lithologic features of the Coast Mountain of the northern Chile. Based on our results, we propose that the coastal relief between Arica and Caldera is fundamentally controlled by structural and lithologic features. Starting from the data of available precipitations, and the latitudinal gradients derived from those, it has not been possible to establish a latitudinal climatic control of the relief.

12 Limitations and Future Work

In terms of the methodology, a general revision of the should be carried out, with the purpose of determining whether the proposed steps are adequate, enough, or new tasks should be added.

In terms of the data, more precipitation data is required, in order to improve the interpolation task and the confidence of the correlation and association models.

In terms of the region under study, we wish to extend the study to other regions. For example, southern Chile where the quantity of precipitations (and the relief) are significantly different than the existent one in the northern Chile. This would allow us to validate and refine the proposed guide, and to contrast the results obtained and described in this paper.

In terms of the evaluation process, we want to incorporate a more formal method of evaluation of results, that complementing the expert domain evaluation, in way of making a more objective interpretation of the results.

In terms of the profiles generated and used in this study, we want to extend the study in order to determine the relevance of the existent relationship among the lithography of an area and the formation of the relief in this area, by considering the latitudinal variations of the precipitations.

References

1. Bacry, E.: LastWave: The signal processing command language. Centre de Mathématiques Appliquees Ecole Polytechnique 91128 Palaiseau Cedex - France, (2004). http://www.cmap.polytechnique.fr/~bacry/LastWave/
2. Brachere, F.: Guimauve: GTK+ Matching Pursuit Decomposition. Observatoire Midi-Pyrenees Laboratoire d'astrophysique de Toulouse, France, (2002). http://webast.ast.obs-mip.fr/people/fbracher/
3. Burbank, D., Anderson, R.: Tectonic Geomorphology, Blackwell Science (2001).
4. Chapra, S., Canale, R.: Métodos numéricos para ingenieros, páginas 503-519, Tercera Edición. McGRAW-HILL,(1999).
5. Craig, C.: Vertical Accuracy Assessment and Archaeological Applications of 90m SRTM DEM in the Norte Chico Region of Coastal Peru, (2005).
6. Crave, A.: Quantification de l'Organisation des réseaux hydrographiques. Thèse Doctorale. Université de Rennes 1. France, (1995).
7. Cuartero, A., Felicísimo, A., Ariza, F.: Accuracy, Reliability, and Depuration of SPOT HRV and Terra ASTER Digital Elevation Models. IEEE Transactions On Geoscience and Remote, Vol 43, No2, February 2005.
8. Dirección General de Aeronáutica Civil, Dirección Meteorológica de Chile, (2005). http://www.meteochile.cl
9. Faundez, P., Fuentes, A., Reumay, P.: Procesamiento Digital de Señales Acusticas utilizando Wavelets. Instituto de Matematicas UACH, (2001).
10. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases, (1996).
11. Glosario Geológico Vademécum REMER, (2005). http://www.proteccioncivil.org/vademecum/vdm02514.htm
12. Gribonval, R., Bacry, E., Abadia, J.: The stft package, Managing short time Fourier transforms and The mp package, Managing Matching Pursuit decompositions. IRISA, Campus de Beaulieu, Rennes Cedex, France, (2004).
13. Hartley, A., Chong, G.: A late Pliocene age for the Atacama Desert: Implications for the desertification of western South America. Geology, 30, 43-46, (2002).
14. Hernández, J., Ramírez, M., Ferri, C.: Introducción a la Minería de Datos. Capítulos 2 y 9. Pearson - Prentice Hall, (2004).
15. Jet Propulsion Laboratory: ASTER Higher-Level Product User Guide. Version 2.0. California Institute of Technology, (2001).
16. Labat, D.: Recent advances in wavelet analyses: Part 1. A review of concepts. Laboratoire des Mécanismes et Transferts en Géologie (LMTG), UMR 5563, CNRS/IRD/UPS, 14 Ave. E. Belin, Toulouse 31400. France, (2005).
17. Lague, D.: Dynamique de l'érosion continentale aux grandes échelles de temps et d'espace: modélisation expérimentale, numérique et théorique. Thèse Doctorale. Université de Rennes 1. France, (2001).
18. Mallat, S., Zhang, Z.: Matching Pursuit With Time-Frequency Dictionaries. Courant Institute of Mathematical Sciencies, Computer Science Departament, New York University, (1993).
19. Modelos digitales de elevación escala 1:50000. Generalidades y especificaciones, (2000). http://mapserver.inegi.gob.mx/geografia/espanol/normatividad /mde/menu.cfm? c=198

20. Montgomery, D., Balco, G., Willet, S.: Climate, tectonics, and the morphology of the Andes. Geological Society of America. v.29, n. 7, p. 579-582, (2001).
21. Quezada, J.: Interacción entre Procesos Endógenos y Exógenos en la Configuración del Relieve del Borde Costero del Anteraco del Norte de Chile. Technical Report. Universidad Católica del Norte. Chile, (2005).
22. Riquelme, R., Martinod, J., Hérail, G., Darrozes, J., Charrier, R.: A geomorphological approach to determining the Neogene to Recent tectonic deformation in the Coastal Cordillera of northern Chile (Atacama), (2002).
23. Unser, M., Aldroubi, A., Eden, M.: A Family of Polynomial Spline Wavelet Transforms. Biomedical Engineering and Instrumentation Program, National Centre for Research Resources. National Institutes of Health, Bethesda. USA, (1993).
24. Unser, M.: Splines: A Perfect Fit for Signal and Image Processing. IEEE Signal Processing Magazine, (1999).
25. Welch, R., Jordan, T., Lang, H., Murakami, H.: ASTER as a Source for Topographic Data en the Late 1990's. IEEE Transactions On Geoscience and Remote, Vol 36, No4, July 1998.
26. Wikipedia, (2005). http://es.wikipedia.org
27. Witten, I., Frank E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, (2005).