Principal Components Analysis (PCA)of Monument Stone Decay by Rainwater: a case study of "Basílica da Estrela"church, Portugal

Carlos Figueiredo[1*], Carlos Alves[2]

1 CERENA - Centro de Recursos Naturais e Ambiente, UID/ECI/04028/2013, DEcivil, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal; carlos.m.figueiredo@ist.utl.pt;

2 LandS/Lab2PT (Landscape, Heritage and Territory Laboratory), UID/AUR/04509/2013; POCI-01-0145- FEDER-007528 and Earth Sciences Department, School of Sciences, University of Minho, Portugal; casaix@dct.uminho.pt

* Correspondence: carlos.m.figueiredo@ist.utl.pt; Tel.: +351-21-841-7023

✔ BASÍLICA DA ESTRELA

• History, architecture and location

- It is the **most relevant 18th century monument in the city of Lisbon,** which **style** with a few baroque elements is classified as being **neo-classical.**
- It is located in **a moderately polluted area** about **15 km** away from **the sea.**
- It was **started** in **1779,** by order of **Queen Mary I,** and **finished eleven years later.**

• Building Materials

- It was built with **Jurassic** and **Cretaceous limestones** exploited at **Lisbon region.** Theses limestones have**:**
  - **several colours (beige, greyish-blue, rose and ochre);**
  - very little **porosity (< 1%);**
  - very low **permeability (from 1.34 x 10^{-1} (mD) to 4.96 x 10^{-1} (mD));**
  - more than **95%** of **calcium carbonate** and less than **3% of silica.**
- The **yellowish variety** is, however, **slightly dolomitic** and **clayey.**

- **Physical weathering** forms such as **granular disintegration, flakes, scales** and **spalling,** prevail **inside.**
- **Chemical weathering** forms are, however**, also** present inside **and are largely dominated by calcite re-precipitation forming large white zones.**

- **Soluble salts** were, *in contrast*, practically **non existent** or **hardly found**.
- **Infiltration** of **rainwater** through the **terrace** of the church was pointed out as being the **main problem** of the monument.

✔ MAJOR GOALS:

- An **extended version** of **Principal Components Analysis** (PCA) approach to monument **stone decay** phenomena is **now presented**.
- This approach will be used, in this paper, to **help data interpretation** and also as a **first step** of a **stepwise approach** to the **eigenvector methods** of data analysis.
- **PCA results** combined with **rainwater sampling** are discussed in the perspective of a **nondestructive tool** (**for** the **characterization** of **alteration** of **geologic materials** in the **built environment)** as it does **not involve** the **extraction** of **samples** from those materials.
- **Only** the **rationale** and the **general methodological procedures** used in PCA will now **be presented**.
- The **mathematics** (theoretical and practical manipulation) and the **computational essentials** underlying PCA implementation, **are beyond** the **scope** of this paper (see Davis, J.C., 1986, for details).

✔ METHODS

❑ *Data Sampling*

- **Physical** and **chemical analyses** were performed on **seventeen seepage water samples** collected over **three years inside** the church at the **elevated choir.**
- Cl^-, NO_3^-, SO_4^{2-}, HCO_3^-, CO_3^{2-}, Na^+, K^+, Ca^{2+}, Mg^{2+}, and **Electrical conductivity** (σ), **pH** and **temperature (T, °C)** were **measured** on **each sample**;
- **Table 1** gives the **basic statistical parameters** of the **raw data set.** This consists of **(the) 12 (previously mentioned variables) physical** and **chemical properties (variables, table columns)** measured on **seventeen** samples.

Table 1. Raw data set: basic statistical parameters of **physical** and **chemical properties** measured on **seventeen seepage water samples.** The **chemical analyses** are in **weight per cent. MAX:** maximum value; **AV:** average (mean); **MIN:** minimum value**; STD:** standard deviation.

	Variables - Physical and chemical properties											
	T (°C)	pH	σ (μS/cm)	HCO_3^- (%)	CO_3^{2-} (%)	Cl^- (%)	NO_3^- (%)	SO_4^{2-} (%)	Na^+ (%)	K^+ (%)	Ca^{2+} (%)	Mg^{2+} (%)
MAX	25.0	11.7	1424.0	310.6	163.7	67.4	20.5	29.9	122.5	330.0	46.5	0.20
AV	19.5	10.3	722.4	111.0	98.5	34.9	9.4	9.1	67.6	177.9	4.4	0.06
MIN	16.7	8.2	100.4	4.0	1.2	23.7	0.00	0.00	41.0	115.0	0.7	0.00
STD	2.3	0.8	305.9	93.2	36.1	11.8	6.0	7.1	22.5	58.2	10.9	0.06

❑ *Principal Components Analysis (PCA)*

- **The data gathered** all over the sampling period **form** the **raw data set** that **was worked out** and **analysed using Principal Components Analysis (PCA),** in this paper. **PCA approach** was used to **help data interpretation.**

• The Rational

- **PCA** is a **factor analysis technique** designed for **interval or ratio data** that are **measurements** made on a **continuous numerical scale (Davis, J. C., 1986).**
- In general, **PCA** as any other **eigenvalue and eigenvector methods** was originally devised to **explain the interrelationships** in a **large numbers of variables** by the presence of **a few factors** or **principal components** or **axes**.
- If the **raw multivariate data matrix** has ***n*** **rows** that represent **observations/samples** and ***m*** **columns of variables**, the *n* **samples** or **objects** may be regarded as being **points** located in the ***m*****-dimensional space** defined by the ***m*** **variables**.
- **PCA** has as **its main purpose** to **decompose** the **larger m-dimensional space** (a **multivariate set of observations**) into a **smaller p-dimensional one,** by *computing* **new, uncorrelated orthogonal components** that are **linear combination** of the **original variables** and **losing** as less as possible of the **variance in the original data set**.
- The **new components** are called ***principal components*** of the **multivariate data matrix**.
- **How many factors** should be retained, is the question now? The **usual assumption** is that **p < m**: we should need **only *p*** **factor** axes to **explain** our data.
- The **linear transformation** of *m* **original variables** to ***p*** **new variables** is performed in a fashion that requires that **each new variable** *accounts for*, successively, **as much of the total variance as possible**.
- **A general pragmatic approach** may consist of **extracting** *only* **two or three principal components** and, then, **plot two** at a time as **2D flat diagrams** that are more **easily manageable** and perceptible dimensions at just one glance.
- **Finally**, the **principal components** have to be **interpreted in terms of original variables**. (However, sometimes this may not be made as easier as we wish).
- **A full circle approach** from **variables** to **principal components**, for **reduction** in the **size** of **the problem**, ***back*** to **variables** for **interpretation** of the **principal components**, is usually used.
- **2D diagrams** of the principal component's **loadings** show the **correlation** among the **original variables** themselves and **also** between **these** and the **principal components** analysis axes.
- On the other hand, **projecting** the **samples scores** (samples co-ordinates) onto the first two principal axes (interpreted in terms of the original variables according to their loadings in the principal components) some significant **insight** into the **inter-samples relationships** in the data set could also be obtained.
- This way, this may **help the analyst** to *explore* the **inter-variable relationships**, the **inter-object** (similitude) **relationships** in a given data set, as well as the **interrelations** between the **variables** and **objects** with **each** of the respective **principal components**.
- As the **principal components** are **linear transformations** of the *m* **original variables**, we are able to plot **PCA scores** simply by **projecting** our **original observations** *onto* the **principal axes**.

• General Methodological Algorithm: a simple layout

- Here, **only** some of the **general or basic computational steps** usually involved in **principal components analysis**, as it is, supposedly, implemented in several commonly **available libraries of computer programs**, will **be pointed out**.
- **PCA** tries to **explore** some of the **mathematical** and **computational relationships** that exist *between* a **data matrix**, its **matrices of cross-products**, and their **eigenvalues** and **eigenvectors**.
- As the **principal components** are nothing more than the **eigenvectors** of a **variance-covariance** or a **correlations matrix**, **PCA** is, then, concerned with *finding* **these axes** and **measuring** their **magnitude**.
- It **starts** by **extracting** the **eigenvalues** and **eigenvectors** of a **variance-covariance** or **correlations matrix**, and then **discarding** the **less important** of *these*.
- The **eigenvectors** are the **coordinates** of the **principal components** axes of the **data set** and may *provide* significant **insights** into the **structure underlying** the *data set*, yielding the **orientations** of the **principal axes** of the **cloud** of **points**.
- The **eigenvalues**, on the other hand, represent the **lengths** of the successive principal axes or **principal components**. That is, they *represent* the **amount** or **proportion** of the **total variance** *transported* or **accounted for** by the **eigenvectors**.
- **In general**, a **PCA** *implementation* may involve **only a few steps**, *starting* by **computing** the **matrix** of **variance-covariance** or **correlations**.
- *Then* it **proceeds** by *extracting* the **eigenvalues** and *their associated* **eigenvectors** from the **matrix** of the **cross-products** of an ***original raw*** or *transformed* data set. (The matrix of the cross-products may be obtained from an original raw or transformed data set. The variance-covariance matrix will contain elements of correlations when all the initial raw variables in the data set are standardized so they have means of 0.0 and variances of 1.0). (*For instance*, **standardization** may be *unavoidable* if the **original variables** are *expressed* in **different**, **incompatible units**).
- In a **third** ***step***, we **compute** what is called **principal component scores** by **projecting** *onto* the **principal components** each **sample** or original observation (Principal components loadings are the elements of the eigenvectors that are used to compute the scores of observations and they are simply the coefficients of the linear equation which the eigenvector defines (*Davis, C. J., 1986).*
- A **final** ***step*** in **PCA** implementation may **involve** the **plotting** and **interpretation** of **2D scatter diagram** defined by each **pair** of the **principal components**. By *cross-plotting*, the **variables** and the **samples** are shown **at positions** *representing*, respectively, their **loadings** and **scores** on the *principal components*. The **arc on the diagram** is part of a **circle** *representing* **a communality** of **1.00**. The **communalities** are the **amount** of **variance** of **each variable** *retained* in the **principal components**. *If a* **variable falls** *on the* **circle**, the **two components** *account for all* of **its variability**. **Variables** that *plot* **inside the circle** are *characterized* by **variability** that is **not represented** *by the two principal components*.

✔ RESULTS AND DISCUSSION

- The **matrix** of **pairwise correlations** and the **eigenvalues** for the *first five* **eigenvectors** are given in **Table 2**.
- The **loadings** of the *eleven* **original variables** on **axis I** are *plotted* against the **loadings** on **axis II**, in **Figure 1a**.
- The **samples** are **plotted** on the **score space** defined also by the **two first principal components** (**Figure 1b**). The **samples** are *shown* **at positions** *corresponding* to **their scores** on the **first two axes**.
- The **first principal axis** *contains* about **45.2 %** of the **total variance**, *whereas* the **second principal component** *represents* an ***additiona*****l 24.7%** (**both** *correspond* to **almost 70.0 %** of the **total variance** of the results from *seepage water samples*).

Table 2. Matrix of pairwise correlations and the eigenvalues for the first five eigenvectors.

	T	pH	σ	HCO_3^-	CO_3^{2-}	Cl^-	NO_3^-	SO_4^{2-}	Na^+	K^+	Ca^{2+}	Mg^{2+}
T	1.00											
pH	-0.35	1.00										
σ	-0.23	0.20	1.00									
HCO_3^-	-0.08	-0.62	0.07	1.00								
CO_3^{2-}	-0.26	0.58	0.24	0.05	1.00							
Cl^-	-0.15	-0.40	0.32	0.90	0.22	1.00						
NO_3^-	0.16	-0.55	0.29	0.72	0.04	0.78	1.00					
SO_4^{2-}	-0.39	-0.01	0.65	0.56	0.41	0.67	0.48	1.00				
Na^+	-0.14	-0.54	0.27	0.87	0.05	0.86	0.81	0.63	1.00			
K^+	-0.10	-0.49	0.31	0.86	0.18	0.94	0.86	0.66	0.90	1.00		
Ca^{2+}	-0.28	0.45	0.58	-0.25	0.07	-0.19	-0.22	0.28	-0.13	-0.23	1.00	
Mg^{2+}	0.22	-0.56	-0.09	0.190	-0.47	0.12	0.42	-0.12	0.46	0.30	-0.24	1.00

Eigenvectors	I	II	III	IV	V
Eigenvalues	5.422	2.960	1.284	0.837	0.599
Percentage of total variance contributed by each eigenvalue					
	45.2 %	24.7 %	10.7 %	7.0 %	5.0 %
Cumulative (%) value of total variance					
	45.2 %	69.90%	80.60%	87.60%	92.60%

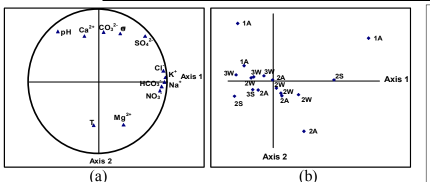

Figure 1. Seepage waters: plot of the **first two principal components** or principal axes. Horizontal axis is principal component I, vertical axis is principal component II: **(a) Variables** are shown at positions corresponding to their **loadings** on the first two axes (adapted from [2]); (b) **Samples** are shown at positions corresponding to their **scores** on the first two axes. In this plot, the **numbers** and the **letters** in the **dot labels** stand, respectively, for the **year of sampling** and the **seasons**: **W**, for **winter**; **A**, for **autumn** and **S**, for **springtime** e (rain is very rare in the summer in Lisbon).

- **K^+, Na^+, Cl^-, SO_4^{2-}, HCO_3^-, NO_3^-** and **pH** are all **well represented** on the **plan** defined by the **first two principal eigenvectors**, with **communality** values varying, respectively, *between* **0.95** and **0.80** (**Figure 1a**).
- *On the other hand*, **Ca^{2+}, CO_3^{2-}, Mg^{2+}, T (°C)** and **specific conductivity** (σ) are **not well represented** on **this plan** as they *plot* **far from the circle** *representing* a **communality** of **1.00.**
- These **last variables** *involving* also the **third** or **fourth** *principal components* **seem** thus to be *less important* to **explain** the **overall variability** in the *seepage water samples*.
- The **first principal eigenvector** is **strongly** and **positively** *correlated* with **K^+, Na^+, Cl^-, HCO_3^-** and **NO_3^-** and **negatively** correlated with **pH**.

- **Samples** *projected* on to **the right side** of the **scores plot** have **values** for these *chemical species* **higher** than their **mean values**. On the **left side** we find the **samples** with the **highest values** of **pH**.
- The **second principal eigenvector** is **positively correlated** with **σ, pH, CO_3^{2-}, Ca^{2+}, SO_4^{2-}** and **negatively** with **T** and **Mg^{2+}.** *However,* **this axis** does **not clearly contribute** for the **analysis** of **sample's position** onto *this* **scores plan** of the two principal components.
- **Only** K^+, Na^+, Cl^-, HCO_3^- and NO_3^- form **one cluster** on this **factor plan** that is, in general, **positively correlated** with SO_4^{2-}. This seems to **suggest** the **same source** or **process** *involving* the **strongly** and **positively** *correlated* **variables** forming the **cluster** *as well as* a **not** very *different* **source** and **alteration process** or *possibly* a *slight* **combination** of **other ones** involving SO_4^{2-}.
- Together **with pH** *these variables* seem to **play a significant role** in the **characterisation** of *seepage waters*. They **explain** *most* of the **variation observed** in the **chemical composition** of the *samples*, while the *other variables*, including **Ca^{2+}**as well, do not. This *surprisingly* **secondary role** *played* by **Ca^{2+}** is *possibly* associated with **stalactite formation** observed **inside** the *church*.
- *However*, **all the variables** *analysed* do **not provide** *enough* **discrimination** of **seepage waters** to *allow* **sub-classifications** *other than* **richer/poor samples** in the *content* of *these variables*.
- *Taking into account* the **sampling period**, *no* **time-** or **seasonal-dependent control** of *seepage water* **composition** has **appeared clearly** *from the analysis of the data*. This could *reflect* a *significant* **uniformity contribution** of **ion sources** and **stone alteration processes**.

✔ CONCLUSIONS

- The **water-rock interaction** and **environmentally-induced processes**, at "**Basílica da Estrela**", seem to **promote** essentially the **enrichment** of *seepage waters* in **K^+, Na^+, Cl^- and HCO_3^-**.
- In this case, **PCA** has *produced* a **result** in **general agreement** with the *one* obtained by the *analysis* of the **same data set** performed by Figueiredo et al. (2000, 2001, 2007) *using* a, perhaps, **more classical geochemical approach**.
- *However*, it **should be stressed** that this **kind of multivariate analysis** *may provide* a **basis** for the **management** of **environmental** and **stone decay data** which may, **when needed**, be **combined** *with other* **geochemical** and **petrophysical studies**.
- For instance, **PCA** *may* **help** in the **planning** of **future searching campaign** for **related studies**. Hence PCA **can help** in **non-destructive studies** of **stone decay** by the *study* of these **samples** that *represent* the **product of interaction** *between* **pollutants** and **stones** *without sampling* the **cultural materials**.

Funding: The authors gratefully acknowledge the support of the CERENA strategic project FCT-UID/ECI/04028/2013) and the LAMPIST of the Instituto Superior Técnico, University of Lisbon as well the Lab2PT, Landscape, Heritage and Territory Laboratory(UID/ECI/04028/2013) of the University of Minho supported by the FCT — Fundação para a Ciência e Tecnologia (Portugal), with Portuguese funds and funds from the European Union (FEDER, Programa Operacional Factores de Competitividade — COMPETE 2020 — Programa Operacional Competitividade e Internacionalização, POCI, POCI-01-0145-FEDER-007528).