**GRAU D’ENGINYERIA INFORMÀTICA (UPC).**

**CURS 20-21 Q2** –**QUIZ 1**

**Anàlisi de Dades i Explotació de la Informació (ADEI) .**

**(Data: 10/11/2020 14:00-15:30 h A5-202 Room)**

**Professor**: Lídia Montero Mercadé

**Rules for quiz:** Internet access is required, emailing and chatting is strictly forbidden. Mobile phones should be switched off.

**Duration:** 1h 30 min

**Marks**: Before 17/11/20 Subject ATENEA WEB site.

**Open Office**: Email requests.

**Problem 1: All questions account for 1 point**

The data set for this exercise contains 396 observations for the mean daily values of the variables included in the table referred to a wastewater treatment plant. The plant has measures on the quality of the wastewater at the entrance of the plant, they are the variables xxx.e, from here they go to a first decantation process (Primary Treatment) where it is intended that they settle the solids in suspensions. Then it goes into the Biological Reactor. This is the most critical part. Here is a biological mud that literally "lives by eating organic matter." It is activated by temperature and aeration. This process is carried out by many species of microorganisms. These microorganisms work at different temperatures. If they are not balanced with the composition of the water, they eat each other. The control variables are those that graduate aeration, temperature, recirculation and purge of the bioreactor. This is the most difficult part: if things go well, dirty water and biological mud enter the bioreactor and end up with clean water and colonies of microorganisms that have consumed the organic matter. Then the water goes through a second decanting process where the microorganisms settle because if the water is not aerated they fall to the ground. And then clean can be poured into the river. The purified water, at the end of the process, before being poured into the river must have neither DBO, nor DQO, nor SS nor SSV, absolute zeros are impossible and therefore the current legislation has permissible limits that are not dangerous for the life in rivers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean Daily Observations | Input | Decantation | Biological | Output |
| Flow (Q) | **q.e** |  | ***qb.b*** |  |
| Iron pretreatment (FE) | **fe.e** |  |  |  |
| Hydrogen potential (pH) | **ph.e** | **ph.d** |  | **ph.s** |
| Solid in Suspension (SS) | **ss.e** | **ss.d** |  | **ss.s** |
| Suspended Volatile Solids (SSV) | **ssv.e** | **ssv.d** |  | **ssv.s** |
| Fraction of degradable organic matter (DQO) | **dqo.e** | **dqo.d** |  | **dqo.s** |
| BIOdegradable organic matter fraction (DBO) | **dbo.e** | **dbo.d** |  | **dbo.s** |
| Volumetric Analysis (V30.B) |  |  | **v30.b** |  |
| Recirculation Flow (QR.G) |  |  | **qr.g** |  |
| Purge Flow (QP.G) |  |  | **qp.g** |  |
| Air inflow (QA.G) |  |  | **qa.g** |  |
| Mixed Liquor Suspended Solids (MLSS.B) |  |  | **mlss.b** |  |
| Volatile solids in suspension liquor mixture (MLVSS.B) |  | | **mlvss.b** |  |
| Cell Age (MCRT.B) |  |  | **mcrt.b** |  |

The list of the included variables in the dataset contains some additional columns that will not be considered in the exercise.

|  |  |
| --- | --- |
| date | id from 1 to number of observations |
| dateformated | dd-mm-yy |
| datenorm | dd/mm/yyyy |
| q.e | Input Flow |
| qb.b | Flow after biological reactor |
| qr.g | Recirculation Flow |
| qp.g | Purge Flow |
| qa.g | Air inflow |
| fe.e | Iron pretreatment |
| ph.e | Hydrogen potential |
| ss.e | Input Solid in Suspension |
| ssv.e | Input Suspended Volatile Solids |
| dqo.e | Input Fraction of degradable organic matter |
| dbo.e | Input BIOdegradable organic matter fraction |
| *nkt.e* | *Input Hydrogen potential* |
| *nh4.e* | *Input Ammonium concentration* |
| *p.e* | *Input Phosphor concentration* |
| ph.d | Decantation Hydrogen potential at the settler |
| ss.d | Decantation Solid in Suspension at the settler |
| ssv.d | Decantation Suspended Volatile Solids at the settler |
| dqo.d | Decantation Fraction of degradable organic matter at the settler |
| dbo.d | Decantation BIOdegradable organic matter fraction at the settler |
| *nkt.d* | *Decantation Hydrogen potential at the settler* |
| *nh4.d* | *Decantation Ammonium concentration at the settler* |
| *p.d* | *Decantation Phosphor concentration at the settler* |
| ph.s | Output Hydrogen potential |
| ss.s | Output Solid in Suspension |
| ssv.s | Output Suspended Volatile Solids |
| dqo.s | Output Fraction of degradable organic matter |
| dbo.s | Output BIOdegradable organic matter fraction |
| *nk.s* | *Unknown* |
| *nh4.s* | *Output Ammonium concentration* |
| *p.s* | *Output Phosphor concentration* |
| v30.b | Biological Volumetric Analysis |
| mlss.b | Biological Mixed Liquor Suspended Solids |
| mlvss.b | Biological Volatile solids in suspension liquor mixture |
| im.b | *Unknown* |
| cm1.b | *Unknown* |
| cm2.b | *Unknown* |
| mcrt.b | Biological Cell Age |
| *trh.c* | *Unknown (non important)* |
| dbo.dqoe | Input Quocient DBO.E into DQO.E |
| dbo.dqod | Quocient DBO.D into DQO.D at the settler |
| dbo.dqos | Output Quocient DBO.S into DQO.S |
| weekday | Day of the week |
| season | Year season |

The data technically correspond to daily measurements and there is a temporal correlation that cannot be dealt with in this subject. **You only have to work in this exercise with the data in randomized order.** The response variables are considered the fraction of biodegradable organic matter DBO.S, degradable organic matter DQO.S or solids in suspension, either volatile (SSV.S) or not (SS.S) in the OUTPUT of the plant. **The response variable DQO.S is initially considered.**

1. Produce a randomized dataset to destroy serial correlation.
2. Missing data have been treated, but some NA coded as 0 values still remain in fe.e and qp.g and have to be treated by applying imputation tools explained in class.
3. Univariant outliers for output variable **DQO.S** are also present and have to be treated. Do it.
4. Are there multivariant outliers? Find them. Try to explain their singularity. Multivariant outliers are not going to be treated in this exercise: keep them as they are.
5. Indicate by using exploratory data analysis tools which are apparently the most associated variables with the response variable (use only the indicated variables). Use also **FactoMineR profiling tools**.
6. Define polytomous factors **f.dbo.s, f.dqo.s, f.sst.s** (from SSV.S plus SS.S) for the covariates according to the legal limit (DBO 25 mg/l O2, DQO 125 mg/l O2 and total suspended solids 35 mg/l). **Profile** **f.dqo.s** factor.
7. A Normalized Principal Component Analysis is addressed using as supplementary variables available factor and xxxx.s output variables. How many axes do you have to retain according to Kaiser criteria? And according to Elbow’s rule? What’s the inertia explained by retained Kaiser-based principal components? Try to explain the meaning of the axes in the first factorial plane. Which are the 3 variables with the greatest correlation with the first factorial plane?
8. A Hierarchical Clustering is addressed. How many clusters are needed to represent 60% of the total inertia.
9. A nondefault criteria for selecting the number of clusters to 3 has to be set. Explain the characteristics of cluster number 3.
10. Use a partition method to group available data into the selected number of clusters found in Question 7. Determine the quality of the partition and plot the resulting partition in the first factorial plane.