Comments of the "Analysis of the Canadian benthic database" presentation planed for the EMRAS II meeting in Vienna

These comments are given as complements of the slides. They can not read independently of them.

Slide2
The slide is just a recall of the context and the available data.

Slides 3 and 4
These slides give information about the work which had already been done by the CNSC. This study is based on a Screening Level Concentration approach, and allowed deriving Lowest Effect Level (LEL) and Severe Effect Level (SEL) concentrations for 9 metals and 3 radionuclides. We consider this method as being an "univariate" one because LELs and SELs are calculated for each contaminant, one by one.

Slide 4
The slide gives explanations about the Screening Level Concentration approach. For a given contaminant (e.g. Uranium) and a given specie (e.g. Acarina), a frequency distribution of the concentrations is done for all the sites where the considered specie is present. The corresponding contaminant concentrations are ordered and the frequency distribution is then drawn. From this distribution the concentration corresponding to the percentile 90%, the so-called SSLC (Species Screening Level Concentration), is determined.

In a second step, all the SSLCs derived previously (one by species) for a given contaminant (still U as example) are ordered and the frequency distribution of the SSLCs is drawn. LELs and SELs are next estimated as the concentrations corresponding respectively to the percentiles 5 and 95%.

Then, if the sediment contamination of a site is below the LEL, environmental risk is likely to be minimal and overruled. If the sediment contamination is above the SEL, there is concern about environmental risk.

Slide 5
The slide presents in few words our approach, which aims to demonstrate whether changes observed in species distributions can be explained by changes in contaminant concentrations in sediments.

Slide 6
Our objective requires building two matrixes, derived from the initial database sent by the CNSC. The first one, the specie matrix, describes the presence (indicated by the value 1) or absence (value 0) of each specie in each sampled site. The second one, the contaminant matrix, contains in accordance with the previous data the concentration values of 12 contaminants on the same sites.

Slide 7
The slide brings information about the species matrix: 196 stations (or sites) are taken into consideration, as well as 209 benthic species. The matrix contains a lot of zero values (absence of species). This is important because zero values lead difficulties into statistical analysis.

Slide 8
The slide brings information about the contaminant matrix. In totality, there are 27% of missing values (no concentration reported), but this percentage varies a lot according to the contaminants, from 0% for nickel to 62.24% for chrome. Finally only 31 sites (stations) are fully informed in terms of contaminant concentrations (no missing values in any contaminant columns).

Slide 9
Thus we can consider two datasets. One corresponds to data (presence/absence values of species and contaminant concentrations) for all the 196 sites, called "all data". The other corresponds to the same kind of data but for the only 31 sites having no missing values in the contaminant matrix; it is called "complete data".

Slide 10
The slide explains our approach.
Firstly we analyzed the "complete data" set using classical methods which permit to describe in which way the species are related to the environmental variables. We used two ordination methods, Redundancy Analysis (RDA) as constrained ordination and Principal Components Analysis (PCA) as unconstrained ordination. Since it is recommended by some authors (Oksanen 2010) we added vectors fitting approach to this last method in order to "explain" the ordination using contaminant knowledge of sites. Objectives and principles of the methods we applied are described in the following slides.

In a second step, as unconstrained and constrained ordinations can't be used with missing values, we tried to develop a method which could be used with such N.A. (Non Available) values and which could permit to bring to light the implication of changes in contaminants on the species distributions. We used this "home made method" with both datasets ("all data" and "complete data" sets).

Slides 11 to 13 give explanations about RDA methods.

Slide 14
To use the RDA method, we first needed to transform the species matrix according to the Hellinger transformation (details about Hellinger transformations are given into the slide 62). According to some publications (Legendre *et al* 2001, Bouchon-Navro *et al* 2005,) RDA after Hellinger transformation is the best constrained ordination method for presence/absence species.

Slide 15
The slide presents the RDA ordination triplot we obtained, and the rules to interpret it.

Slide 16
Actually, it is better to restrain the number of constraints (contaminants variables) when a RDA is done. So, as second step, we had to select the most significant constraints. Having no idea about which ones are important, we used an automatic procedure. Starting from a model having no variable (null model), the procedure adds successively each contaminant variable and estimates at each step the Akaike criterion ("AIC", based on likelihood). Lower is the AIC, better is the model. The table on the left shows that how substances are classified on this base, considering individually each contaminant. Starting from the null model, adding copper alone is the best solution. Adding vanadium alone leads to the second better model, and so on. The table on the right shows the second step results: the procedure starts with the model based on copper, and adds another variable. The best model for two variables is then obtained when coupling Copper and Vanadium.

Slide 17
Finally the optimal model contains 3 significant variables (according a significant level at 5% - not shown in the slides): copper, vanadium and chrome. Some information concerning the RDA is given (see the slide)

Slide 18
According to the rules of RDA interpretation presented in slide 15, we tried to summarize which species are positively and negatively correlated with the 3 most significant contaminant variables.

Slide 19
On the RDA ordination graph some sites are very close. Sites showed in the example are geographically correlated, since they are situated in a close area. Thus we wondered if there may be a confusing variable, i.e if the distributions of the species could also be controlled by another kind of environmental variables.

Slide 20
The slide presents the conclusion about RDA approach applied to the complete dataset.

Slide 21
The slide gives some information about Principal Component Analysis.

Slide 22
To use the PCA method, as well to use the RDA method, we needed at first to transform the species matrix according the Hellinger transformation (details about Hellinger transformations are given in the slide 62).

Slide 23
The slide presents the PCA ordination biplot we obtained and the rules to interpret it.

Slide 24
The slide gives some information about the PCA results we obtained. The two first Principal Components contribute for 24% of the total variation (of the data). It seems that in the field of ecological communities' analysis it's not a so bad result. We wondered if the principal components (or axes of the biplot), which are linear combination of species, have a meaning. Perhaps species with a high score in the first axis could have specific environmental or communities features. The same question is valid also for the second axis. To answer that question we would need benthos expert's advice.

Slide 25
The slide explains that when an unconstrained ordination method is used, in a second step, it's possible (and recommended) to fit the environmental (or contaminant) vectors onto the previously obtained PCA biplot. This approach permits to "explain" the obtained ordination (sites and species) using ecological (here contaminants) knowledge on studied sites.

For that, a least square regression is done with concentrations as response variable and with scores on the principal components as explanatory variables. The tool we used (vegan package for R software) gives several results. For each contaminant, it gives the coordinates of the vector on the two first principal components. Then the square correlation coefficient between ordination biplot and environmental variable (r^2) is given. Finally, a statistical test is made using permutations, in order to assess if r^2 is statistically significant; the p-value of this test is given in the last column. The presented results were partially cut. They are displayed in totality into the following slide.

Slide 26
The slide displays the PCA ordination biplot with all the contaminant vectors fitted onto (the significant and not significant contaminant vectors, according to the results presented in the left part of the slide).

There are two main groups of vectors : one containing uranium, nickel, lead, ^{226}Ra, ^{210}Pb , ^{210}Po and Copper (colored in green) ; the other containing arsenic, moly and selenium (colored in pink). There are also 2 isolated vectors: vanadium and chrome.

Slide 27
The slide gives the rule of interpretation of the vector fitting approach. In the second point we explain that the interpretation of the species location in regard to the environmental vectors is not clear yet. The angle between a specie and an explanatory variable (contaminant) in the biplot seems to reflects their correlation (need to be checked).

Slide 28
The slide displays the PCA ordination biplot but with only the significant contaminant vectors fitted onto (according a significant level of 5%, cf. red stars in slide 26).

Slides 29 to 31
These slides are a comparison of the results obtained respectively using the RDA and PCA methods.

Slide 29 concerns the sites, slide 30 concerns the contaminants and the last slide concerns the species.

Slide 32
This slide concludes about RDA vs PCA results. The point 3 means that these two methods permit to assess which contaminants have a significant effect onto the benthos community distribution.

Slides 33 to 42
Concern the method we developed and its application to the "all data" set. This method consists in 4 steps: the first is a PCA, the second is the building of site groups, the third is a characterization of these groups in terms of contaminant concentrations and the fourth is the characterization of these groups in terms of species.

Slide 33
The slide explains that, as for the "complete data" set study, in order to use the PCA method on the species matrix, we first applied the Hellinger transformation.

Slide 34
The slide is just a reminder of the PCA biplot's rules of interpretation (same than previously).

Slide 35
The slide displays the contribution of the variance of the two first principal components of the PCA. It's a little bit less than the result obtained with the PCA applied to the "complete data" set. It the second point we wonder if the principal components (or axes of the biplot), which are linear combinations of species, have a meaning. Perhaps species with a high score in the first axis have specific environmental or communities features. The same comment applies to the second axis. To answer that question we need benthos expert's advice.

In the third point, we introduce the fact that the sites seem graphically being distributed in 3 groups. Thus, in the second step of this method we splitted the sites into 3 groups using a specific classification method.

Slide 36
The slide displays the groups of sites we obtained using *kmeans* approach. More details about this method are given in the slide 63.

Slide 37

In the third step of the method we developed, we estimated the mean of each contaminant concentration for each of the three groups. Confidence intervals are estimated using bootstrap approach. More details about bootstrap are given into slide 64.

The Slide 37 is a graphical representation of these means and confidence intervals for 8 contaminants. Using the confidence interval overlap approach we can assume that there is a significant difference in uranium means between groups 1 and 2 and group 3. It's seems to be the only one statistically "significant" difference.

Slide 38

The Slide 38 is a graphical representation of these means and confidence intervals for the 4 remaining contaminants. Using the same approach we can assume that there is a significant difference in selenium mean concentrations between groups 1 and 2. Same conclusions with the mean concentrations of ^{210}Pb and ^{210}Po between groups 1 and 3.

Slide 39

The slide displays the main contaminant concentrations characterizing each group of sites. The sign + means that the concentration of the considered contaminant is high (vs the mean in the other groups) ; the sign - means that the concentration of the considered contaminant is low (vs the mean in the other groups). According to the information displayed in slides 37 and 38, the group 1, for example, is characterized by high concentrations in uranium, ^{210}Pb and ^{210}Po.

Slide 40

The slide displays the main species characterizing each group of sites. For example the specie called "chaoborus" is present in 45.6% of the group 1's sites (high right corner).

Slide 41

The slide displays together the characterization of each group of sites in terms of contaminant concentrations and presence of species. It permits to have a view about the combination "contaminant-species" which characterize each of the three groups of sites.

Slides 42 and 43

The developed method presented in the previous slide was also applied to the "complete data" set (previously treated with constrained and unconstrained ordination methods). As with the "all data" set, 3 groups of sites have been constituted using *kmeans* method. Hence, slides 42 and 43 display the mean contaminant concentrations with their confidence interval.

Slide 44

As the slide 41, which concerns the "all data" set, the slide 44 displays together the characterization of each site group in terms of contaminant concentrations and presence of species for the "complete data" set. It permits to have a view about the combination "contaminant-species" which characterizes each of the three groups of sites.

Slides 45 to 53

These slides display a comparison of the "contaminants-species" pattern observed when "all data" and "complete data" sets are analyzed with the method we developed.

The comparisons can be made only on the basis of contaminants identical and to similar levels. The only ones present contaminants in high and low concentration levels in groups belonging to both studied datasets are ^{210}Po and ^{210}Pb. Thus the only comparisons we could do are between groups having high concentration level in ^{210}Po and ^{210}Pb in both datasets (e.g. groups 3 vs A – in light blue) and between groups having low concentration levels in ^{210}Po and ^{210}Pb in both datasets (e.g. 1vs C - in purple).It must be noticed that, moreover, these groups are characterized by the presence of other different contaminants (i.e. group 3 is also characterized by high concentration in uranium, and group A is also characterized by high concentration in lead, copper and a low concentration in moly)

Nevertheless some same combinations can be observed. In both datasets, Procladius is the most present species in sites belonging to groups characterized by high concentration in ^{210}Pb and ^{210}Po (25.4% of sites in "all data" set and 30.25% of sites in "complete data" set)

On the other side, for both datasets, Chironomus, Chaoborus and Pisidium are members of the 5 most frequent species in groups of sites characterized by a low concentration level in ^{210}Pb and ^{210}Po

However, in these last groups of sites (characterized by low concentrations in ^{210}Pb and ^{210}Po) the Procladius specie is also relatively frequent (in 21.5% of the sites from the "all data"set and in 20.12% of the sites from the "complete data" set) . As the same, in the group of sites characterized by high concentration in ^{210}Pb and ^{210}Po of "all data" set, there is a moderate percentage of sites where the species Chironomus is observed.

Slide 54

The slide displays a summary and comments of the results obtained using the method we developed.

Since analyzing "alldata" and "complete data" sets some same "contaminants-species" combinations appear, it seems that the method we developed is able to bring to light real "contaminant-species" combinations. And then, that the method we developed is reliable to bring some support for the hypothesis of contaminant control on the benthos distribution. Nevertheless, as we have seen in the previous slides some species (e.g. Procaldius) are present in both groups of sites characterized by high and low concentration in ^{210}Pb and ^{210}Po. It probably means that our method misses precision.

Slide 55

The slide explains that the classical method of ordination (PCA+vectors fitting or RDA) and the method we developed are actually two different perceptions, and thus it's quite difficult to compare each other.

Slide 56
The slide explains that despite their difference of perception, when a contaminant turns out to have a big influence according to the PCA method and when its coordinates on the biplot place it in the middle of a group of sites put in evidence by the developed method, then this last one is also able of bringing to light it influence(i.e. lead, copper).

Slide 57
The slide is strength and weakness' summarize of the method we developed and of the classical ordination methods.

Slides 58 to 60
The slide describes the perspectives of the presented work.