Implementation of covariate studies in an Epidemiological G.I.S. *

Virgilio Gómez Rubio, Juan Ferrándiz Ferragud, Antonio López Quílez
virgilio.gomez@uv.es, juan.ferrandiz@uv.es, antonio.lopez@uv.es

2nd December 2002

Abstract

When performing an Exposure Analysis with the R.I.F. the user can define some bands around one or several pollution sources and calculate the rates and relative risks for each band. How distance is defined is crucial. The current approach always supposes a spatial distance when grouping zones into bands.

We propose the use of another grouping method using covariates. Zones are grouped according to the value of a given covariate. This method allows us to evaluate the relationship between covariates and diseases.

Contents

*Appendix to the Final Report of the second period (2002) of the Euroheis Project, submitted to the *European Commission, Directorate-General Health and Consumer Protection* under the title *Covariate Study*.

1 Introduction

The Rapid Inquiry Facility (R.I.F.) is an Epidemiological and Geographical Information System for disease mapping and risk assessment which is being developed within the *EUROHEIS* project, funded by the European Commission.

The actual prototype of the R.I.F. allows the user to perform Exposure Analysis around a pollution source, which can be a point (for example, a nuclear plant), a line (a road or a river) or a region (a lake). The zones nearby are grouped into several bands according to their distance to the pollutant source. After that, rates and relative risks are calculated for each band, so we can compare one another to detect whether there is a higher risk in the inner bands.

In Spanish case-studies we need to group municipalities according to the value of several covariates rather than using spatial distance to a given pollution source.

For example, take the case study of finding out whether there is statistical relationship between drinking water quality (through the measurement of calcium, magnesium and nitrates) and cardiovascular mortality from cardiovascular in Valencia.

Grouping zones by covariate levels is conceptually very similar to grouping by distance, since we can think that the value of a given pollutant is higher in the inner bands and lower in the outer ones. The relevant thing is the exposition level rather than whether the zones are close together or far apart.

As much as smoothing of SMR's is based on exchangeable models (not depending on geographical neighbours but on the whole set of bands) the estimation procedure of relative risks remains the same as in the current *Exposure Analysis*.

The aim of this paper is to describe a new type of study proposed by the Spanish Partner, based on the use of covariates which can be used as a new method for grouping and a new stratum for standardisation.

In the next section we describe how exposure analysis with a single covariate is performed. In Section 3 implementation details are given, where modifications made in ArcView scripts and Oracle databases are explained. Section 4 is dedicated to introduce how more than one covariate could be used and how to standardise using a new stratum based on covariates. Finally, Section 5 describes more deeply how to work with more than one covariate, how to use covariates for standardisation and implementation details about these features.

Throughout this document screenshots we have placed figures shows screenshots of the R.I.F. user interface. These pictures are always related to the surrounding text.

2 How the analysis is performed

We have added a new option to the list of available types of analysis. In the *Study Details* dialog box, clicking on the button labelled *Covariate mapping* a new dialog box will appear. The user can select there the region of study (like in Disease Mapping) and the covariate to group the zones. This is done choosing

- the table where the data are stored,
- the field which contains our covariate,
- the year,
- and how the covariate must be split

in order to generate the bands.

Only numerical covariates can currently be used. It is not important whether they are discrete or continuous. Once the user has defined several intervals (from the minimum value to the maximum) the R.I.F. will group into the same band all the zones whose values of the covariate belong to the same interval.

To define the intervals the user must only write the limits of the intervals (that we have called *breaks*) in a text box. For example, if the *breaks* are 10, 20 30, the intervals used when grouping are [min, 10], (10,20], (20, 30] and (30, max], corresponding to bands 1, 2, 3 and 4. Band number 0 is reserved for those zones where the covariate was not measured or we have no data

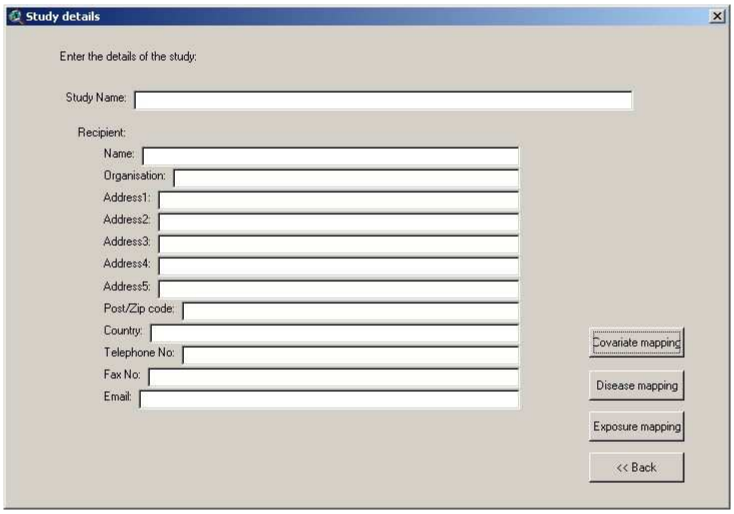

Figure 1: A new button has been added to select a Covariate Study

available. This band will not be considered in posterior analyses (i.e., no rates and relative risks will be calculated).

It is clear that the choice of the breaks is crucial in the analysis, so, before setting them, it is possible to draw a map in order to preview the geographical distribution of the covariate. By default, a quantile legend is chosen (i.e., the intervals are given by the quantiles), but the user can also provide their own breaks for the intervals.

Note that, once the user has selected the study region and covariate, the comparison region and the investigation options are chosen in the same way than before. The final map and reports are also displayed as usual.

Once the zones are grouped, the statistical methods are the same than in *Exposure Analysis*, because we are also working with bands.

3 Implementation

3.1 New tables

In order to handle all the tables related to covariates, a new table has been added to the R.I.F. Its name is *rif_covariates*, and has four fields:

- `tablename` (NOT NULL VARCHAR2(30))

  The table where all the data related to a given covariate are stored.

- `covname` (VARCHAR2(120))

  The name of the covariate itself. This name is displayed to the user when selecting the covariate, through a list.

- `units` (VARCHAR2(40))

  The units in which the covariate is measured. This field can be left blank but it is displayed in the reports and legends.

- `comments` (VARCHAR2(256))

  Some comments about the data (for example, provider or annotations about the quality of the data). It can be blank too.

In the Graphic User Interface of the RIF, the list displayed to pick the covariate will be automatically filled with the values of the field `covname`. Once the user selects the covariates, two new lists will be filled: one with the available years and the other with the fields of the table `tablename`.

Each covariate is stored in a separate table. This table must have at least three fields:

- `join field`

  It must be the same fields that those used in the Mapping Units list that appear in the Covariate Study dialog box, since it is used when joining this table to the data in the shapefiles.

- `year` (NOT NULL NUMBER(4))

  The year the data was measured. This can be changed easily to support full dates.

- `data fields`

  These fields store the covariate data and its name can be any one. It is possible to have several fields in the same table but, for convenience, all must be related to the covariate. For example, we have a covariate called *Students*, and we have separate fields for Primary School, Secondary School, etc.

When the covariate is selected, it is joined to the tables provided by the shapefiles according to the `join field` and before bands are built up.

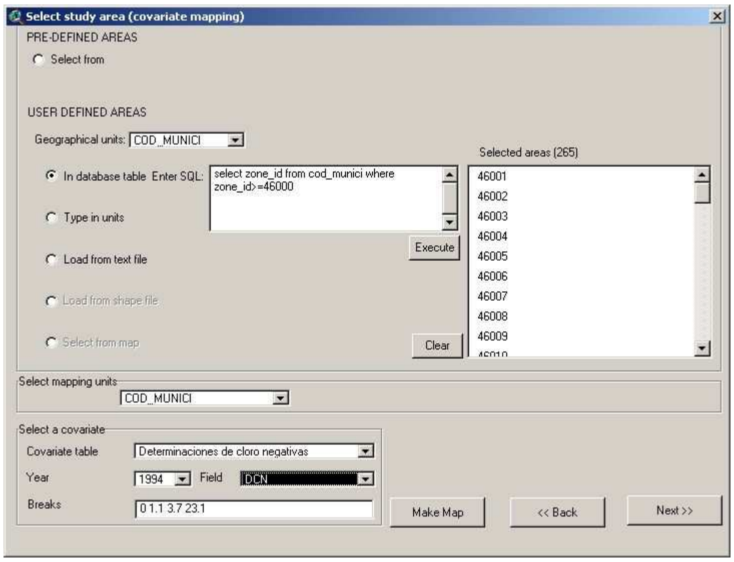

Figure 2: Dialog box used to select study zones and covariate

3.2 New dialog box

A new dialog box has been added to the R.I.F. It is basically the Disease Mapping dialog box but several controls have been added in order to allow the user to select the covariate to use.

In this dialog box we can set the study zones and the covariate for grouping (table, field and year). Once this is done, the user can draw a map by clicking on the button labelled *Make Map*.

3.3 New scripts

Several scripts have been added in order to make the calculus needed. Other old scripts have been modified to add some lines of code to check whether we are performing a Covariate Study.

The names of the new scripts are derived from the names of the old scripts. All their names begin with the word *Covariate*, which used to be *Exposure* or *Disease* in the old scripts. Of course, the task of a new script is analogous to the task of the old script.

For example, the script *MakeBandsTable* is called in an Exposure Analysis and it performs the grouping of zones into bands. The script that builds the bands in a Covariate Study is called *CovariateMakeBandsTable*.

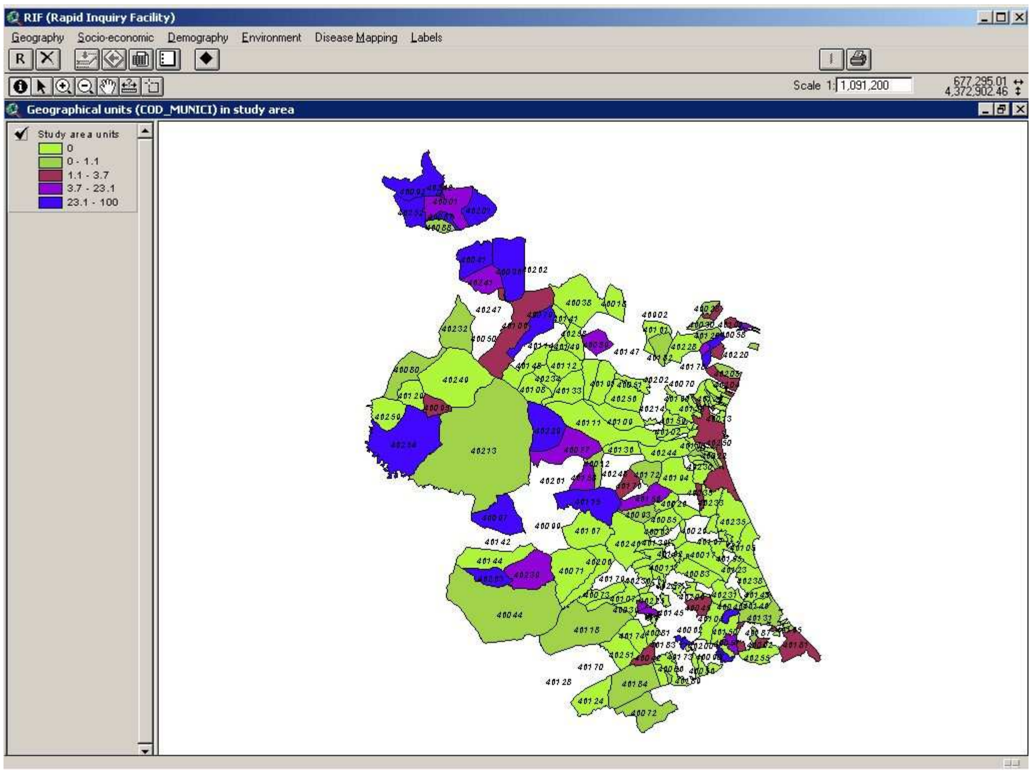

Figure 3: Map generated using the *Make Map* button

The list of modified old scripts is: *InvDoStudy*, *ConvertListGeogUnits*, *OpenCompRegion*, *CloseCompRegionDialog*.

3.4 Final Map

After the end of the query to the Oracle database to calculate the Relative Risks and the Rates, a map is displayed. It has two themes: a background coverage, which shows all the zones, and other theme which represents the study region and the bands built from the covariate.

3.5 Forms and Reports

The reports used in a Covariate Study are the same than those used in Exposure Analysis. Since a Covariate Study is coded using a new identifier, we have added a couple of lines to the Form *rifform*. The first line is just to

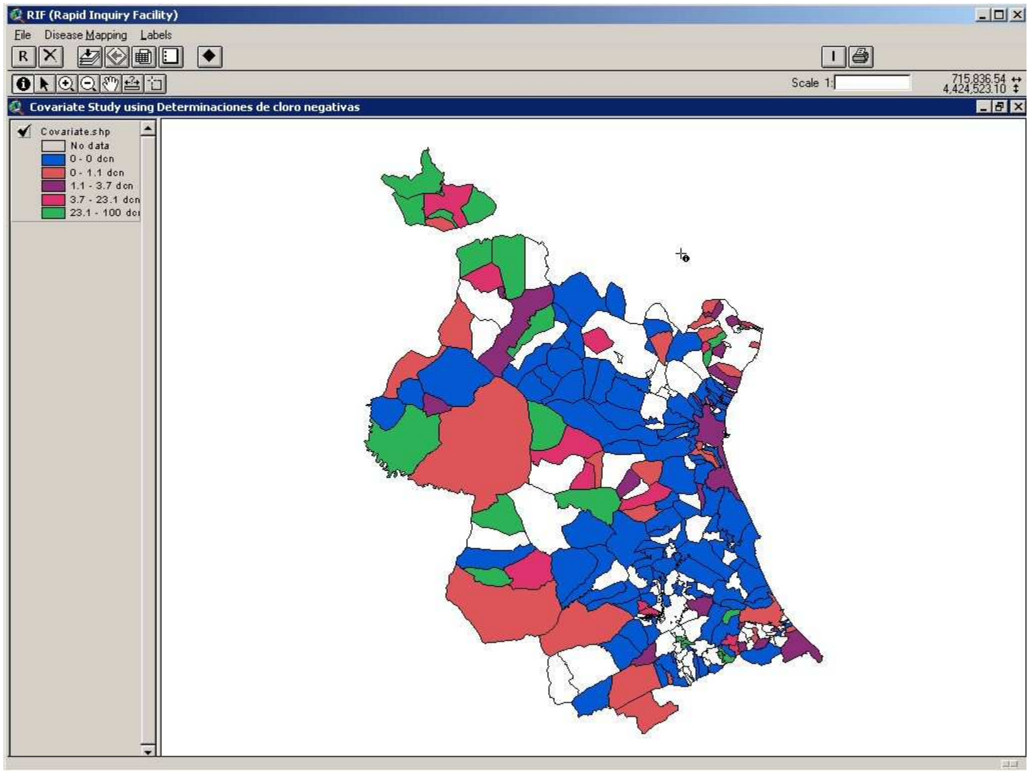

Figure 4: Final map showing the grouped zones

know that we are performing a Covariate Study and the second is a call to the forms of the Exposure Analysis.

4 Further stuff about using covariates

4.1 Using more than one covariate

The use of several covariates can be handled with these new tools, but it requires previous work. We must build a new index that represents, for each zone, the group it belongs according to the list of possible combinations of levels of covariates. For example, if we have three covariates, and a zone has its values belonging to groups 1, 2 and 1, the value of the new index would be 121. Later, this labels will be recoded from 1 to the total number of groups.

We can treat this index as a new single covariate and use the R.I.F to perform the study in the way we have explained before. The breaks can be set easily to get the grouping desired.


rifreportexposureindirect: Visor

Fichero Ver Ayuda

Página: 1

(Ages: 0-85+, Years: 1987-1990, DEPRIV_INDEX: 0-0)

0 - 0 dcn

SDI standardised	Sex	Observed	Expected	RR	Lower 95% CI	Upper 95% CI
No	Males	753	808.53	.93	.87	1
No	Females	527	526.21	1	.92	1.09
No	Males+Females	1280	1334.75	.96	.91	1.01
Yes	Males	753	808.53	.93	.87	1
Yes	Females	527	526.21	1	.92	1.09
Yes	Males+Females	1280	1334.75	.96	.91	1.01

0 - 1.1 dcn

SDI standardised	Sex	Observed	Expected	RR	Lower 95% CI	Upper 95% CI
No	Males	1442	1517.93	.95	.9	1
No	Females	1088	990.12	1.1	1.04	1.17
No	Males+Females	2530	2508.05	1.01	.97	1.05
Yes	Males	1442	1517.93	.95	.9	1
Yes	Females	1088	990.12	1.1	1.04	1.17
Yes	Males+Females	2530	2508.05	1.01	.97	1.05

1.1 - 3.7 dcn

SDI standardised	Sex	Observed	Expected	RR	Lower 95% CI	Upper 95% CI
No	Males	353	390.77	.9	.81	1
No	Females	232	262.29	.88	.78	1.01
No	Males+Females	585	653.05	.9	.83	.97
Yes	Males	353	390.77	.9	.81	1
Yes	Females	232	262.29	.88	.78	1.01
Yes	Males+Females	585	653.05	.9	.83	.97

Figure 5: Report showing the Relative Risks in each band

In the next section it is described the way several covariates can be used and the new dialog boxed that have been added.

4.2 Filtering the effect of covariates

In case we were interested in filtering the effect of a covariate, we could also stratify the population by this covariate, in the same way we do with age and sex. When performing a study, this covariate may be handled as we do with any *Deprivation Index*. This will require very little changes in the source code of the R.I.F.

The expected number of cases in zone i will be

$$E_i = \sum_j r_j N_{ij}\,, \quad j = (a, x, c) \begin{cases} a = 1, \ldots, \text{Number of age groups} \\ x = 1, \ldots, \text{Number of sex groups(?)} \\ c = 1, \ldots, \text{Number of covariate groups} \end{cases}$$

where the index j is a composite index for all the combinations of age-sex groups and levels of the covariate, r_j stands for the specific rate of stratum j and N_{ij} represents the population of stratum j in zone i.

Covariates associated to zones have a single value in each one. All people in every zone will share the same covariate value. Nevertheless, the computational procedure used to filter a *Deprivation Index* (to compute SMRs) can also be applied here if we assign the value of each zone to all of its cases.

5 Using several covariates

5.1 Introduction

This part can be considered an addendum to the previous sections, since here it is described what have been done in order to provide the R.I.F. with tools to use more than one covariate in the analysis.

The approach followed is quite simple. The user can select up to four covariates in the same way than before (i.e., covariate, year, field and breaks). Bands are built by taking all the possible combinations of the subgroups defined for each covariate.

Besides, these bands can be used to filter the effect of the covariates in the same way that a deprivation is used, as it was suggested before.

5.2 Implementation

A new button has been added to the *Covariate Study* Dialog Box, which opens a new window called *Covariate Grouping.* This button has the label *Use more than one covariate* (see Figure 6).

The new Dialog Box displays four control panels which the user can use to select the covariates that will be used. The user can select up to four covariates, together with the years, fields and breaks to use. Everything is done in the same way that with one covariate, as it was explained before.

When all the covariates are selected, the user has to select how many of them will be used. This is done in the Combo Box which is labelled *Select the number of covariates* (see Figure7).

Finally, the user must click on one of these two buttons:

- *Create table for filtering*

  This is the option to use the groups to filter the effect of the covariate. A new index is built taking the same value of the band for every study zone and a table called *COV_SINDEX* is populated with these values. This table is described below.

- *Exp. Analysis with Covariates*

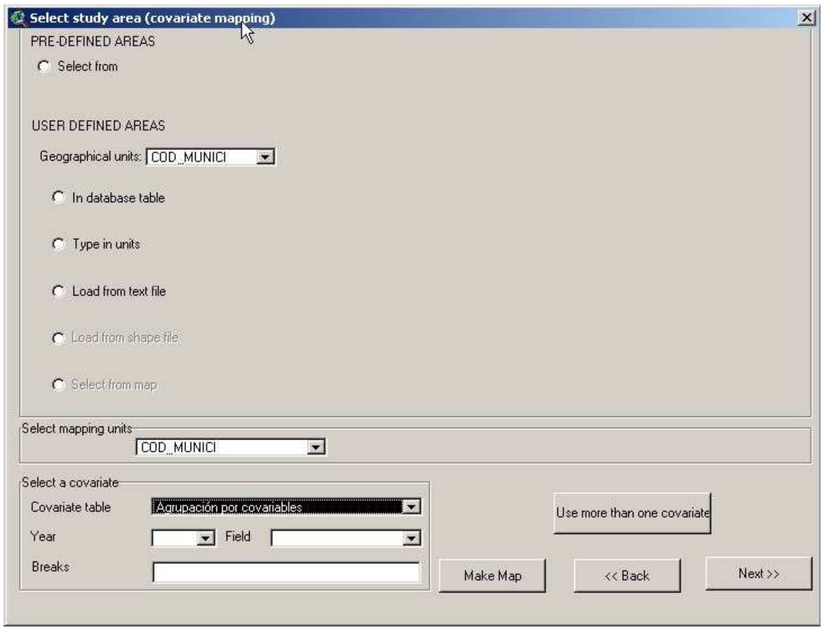

Figure 6: A new button has been added to the Covariate Dialog Box to use more than one covariate.

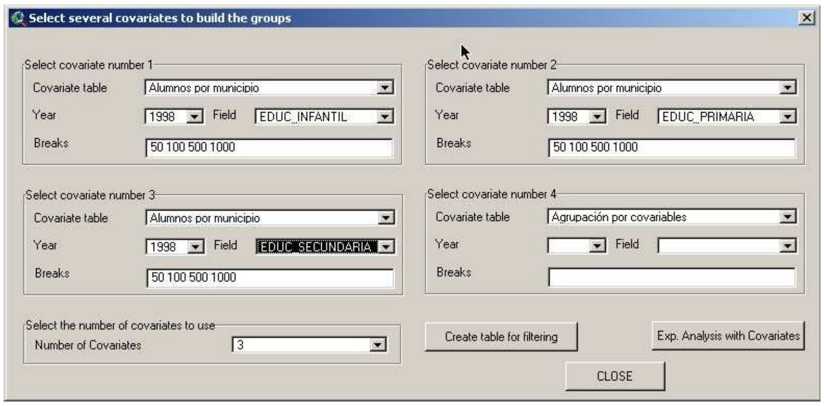

Figure 7: Window used to select several covariates.

Select this option to perform an Exposure Analysis grouping by the bands built from the covariates. All the groups created and their related locations are stored in table *COV_GROUPS*.

The analysis is the same that we did with one covariate, but now we can use more than one.

To return to the previous dialog box there exists another button (labelled *Close*).

5.3 Filtering the effect of the covariates

As we said before, this option is useful when the user wants to take away the effect of the covariates. The R.I.F. creates a new index equal to the groups related to the covariates, so this index can be selected from the list of *Deprivation Indexes* in the *Investigation* Dialog Box (see Figure 8).

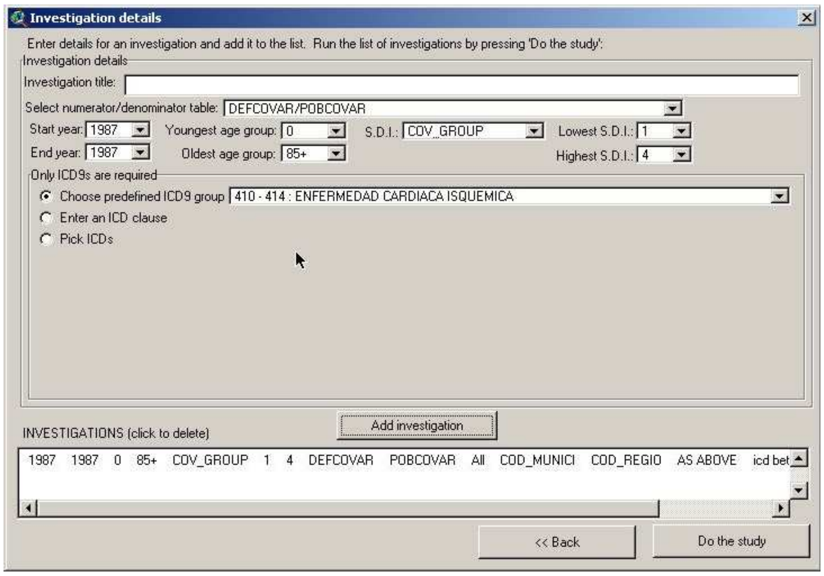

Figure 8: Example of covariate study using the Covariate Index.

The indexes and their related zones are stored in a table called *COV_SINDEX*. This table has several fields, basically the geographical units and a column called *COV_GROUP* to store the index value related to that zone. For our data, these are the fields of this table [1]:

- COD_MUNICI

  This field store the codes of the municipalities, our smaller geographical unit.

- COD_PROV

  Stores the code of the province where the municipality is placed.

[1]Notice that the geographical fields must be changed in order to be able to use the index with other geographical units

- COD_REGIO

  The region where the municipality is placed.

- YEAR

  This field is not really used, but it has been created for convenience. Maybe this can be used later, to create several indexes for different years.

- COV_GROUP

  This is the group to which the municipality, province and/or region belongs to. It is calculated from the covariates and their breaks. More information about how the group is calculated can be found below.

The Combo Box to select the mapping units of a Covariate Study shows the units for which the levels of the covariates will be calculated. For our studies, they are fixed to *COD_MUNICI*. This means that all the groups are related to a municipality when they are calculated and that all the other geographical fields will not be filled and used. This is done because we are interested in studying diseases for the lowest level of aggregation.

With this new index, we have to standardise the population according to it. Since municipality is the lowest level of aggregation, we will sum the population by province, year, age/sex group and covariate index (we deprecate the deprivation index here). We will repeat this steps to calculate the population for every region, year, age/sex group and covariate index.

The mechanism is the same that we would use for a deprivation index.

These populations must be stored somewhere. For municipalities, just a new view is needed. This view (called *POBCOVAR_COD_MUNICI*) joins our main population table and COV_SINDEX by the values of COD_MUNICI. For provinces and regions two new tables are needed, since several sums are performed. These tables are called *POBCOVAR_COD_PROV* and *POBCOVAR_COD_REGIO*.

The results are also linked to numerator tables through views. No table has to be created or filled. For our data, the new views are:

- *DEFCOVAR*, from *DEFUNCIONES*.
- *MORBCOVAR*, from *MORBILIDAD*
- *MORBSINREPCOVAR*, from *MORBILIDAD_SIN_REP*.

In case we wanted to use other geographical unit other than COD_MUNICI, some code must be rewritten. We will need a view for the main mapping units and some tables for the others.

5.4 Exp. Analysis with Covariates

Now, the groups created from the covariates are used like *bands* in a Exposure Analysis. The results are stored in a table called *COV_GROUPS* which has the same structure than the table *COV_SINDEX* described before.

This new table can be considered as a covariate and it can be selected from the *Covariate Study* Dialog Box, as we did before for single covariates. A new entry must be added to the table *RIF_COVARIATES* to handle this new table. In our case, the name of this new *covariate* is *Agrupación por covariables (Grouping by ciovariates)*.

When the user clicks on the button mentioned before, the table *COV_GROUPS* is filled, and the breaks are set in the *Covariate Study* Dialog Box. These are integers from one to the number of groups minus one. This assures that all the bands are taken separately, since the value of the covariate are grouped taking all the values greater than i and lower or equal to $i+1$. The maximum is added automatically to the list of breaks, so this band is also taken.

5.5 Maps and reports

The maps and reports work in the same way than for an Exposure Analysis. Of course, the municipalities in the same band don't have to be close one another, but they can be far apart (see Figure 9).

5.6 Additional changes needed

Minor changes need to be done in the the PL/SQL script called *RIF_QUERY*. In the *if* sentence, where the denominator table is tested to see if we are using the population table called *POPULATION_UK*, this name have to be changed to the name of the populations used. For our data, we use two tables: *POBLACION* (i.e., population) and *POBCOVAR* (i. e., population with the covariate index).

So, the *if* clause we use is

```
if denom_tab='POBLACION' OR denom_tab='POBCOVAR'
```

instead of

```
if denom_tab='POPULATION_UK'
```

5.7 How the bands are calculated

The better way to explain it is through an example. Let us suppose we use 3 covariates and that we split their values into 3, 2 and 2 intervals, respectively. We will label the intervals, for each covariate, from zero.

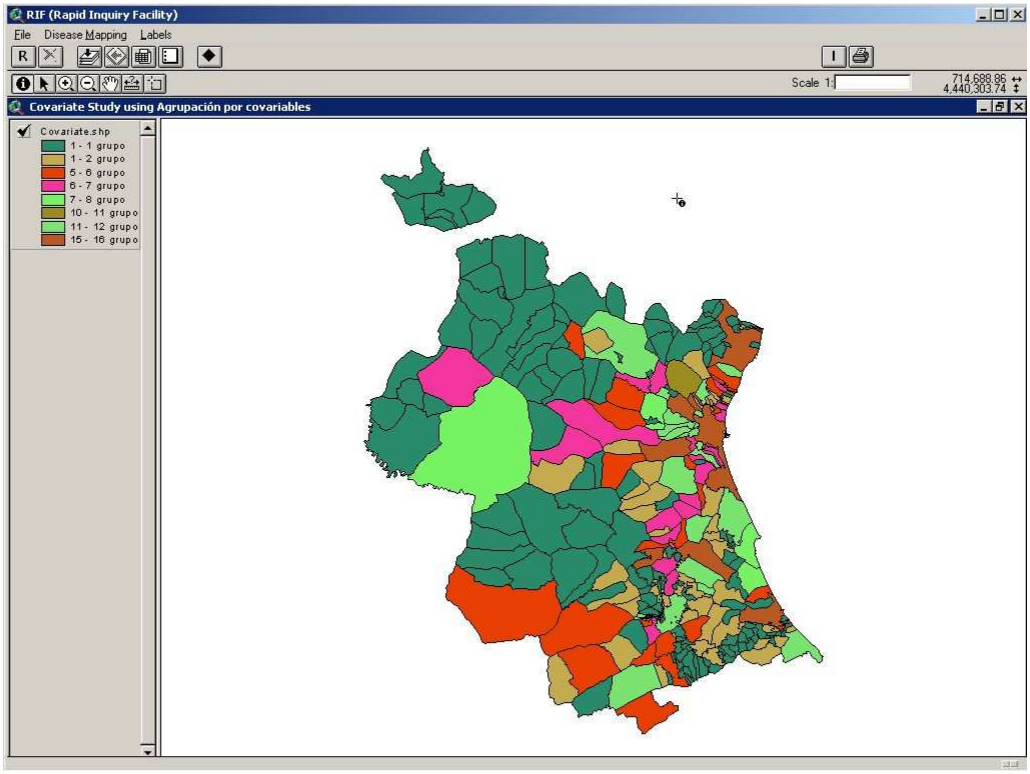

Figure 9: Map generated in a Covariate Study using several covariates.

So, the first group is built taking all the regions whose values of the covariates lay on the intervals number 0, 0 and 0. The second group will correspond to the regions whose values lay on the intervals 0, 0, 1. And so on.

This table will resume all the groups:

Group	1st Cov.	2nd Cov.	3rd Cov.	Group	1st Cov.	2nd Cov.	3rd Cov.
1	0	0	0	7	1	1	0
2	0	0	1	8	1	1	1
3	0	1	0	9	2	0	0
4	0	1	1	10	2	0	1
5	1	0	0	11	2	1	0
6	1	0	1	12	2	1	1

Figure 10: Example of how band groups are built.

In the map legend, the labels we found follow the pattern '*from* - *to* group', which means that the colour shown belongs to the bands in the interval (*from*, *to*]. Since band numbers are $1, 2, \ldots, n$ this means that label (i, i+*1*] only represents band number i. The first band is an exception to this rule and it is represented by '1 - 1 group'.

Although this procedure may appear tricky, it allows compatibility with previous procedures, data structures and functions. Creating custom labels would require major modifications in the source code.