Biometrical Letters
Vol. 46 (2009), No. 1, 15-25

Relating the incidences of AIDS and opportunistic diseases in the European Union

Inês Jorge Sequeira[1], João Tiago Mexia[1], Sandra Nunes[2]

[1]Department of Mathematics, Faculty of Science and Technology, New University of Lisbon,
Quinta da Torre 2829-516 Caparica, Portugal, e-mail: ijs@fct.unl.pt
[2]Department of Mathematics, EST/IPS, Campus do IPS, Estefanilha, 2910-761 Setúbal, Portugal,
e-mail: snunes@est.ips.pt

SUMMARY

Our study was focused on the relation between AIDS and opportunistic diseases. We aimed at obtaining a geometrical representation for those EU countries for which there was sufficient data. A linear representation was obtained, with Portugal at one extreme, followed by the south European countries and lastly the north and central European countries. This representation was validated through an "experiment" in which the population of Luxembourg was increased first 50 times and then 100 times.

Key words: AIDS incidence, opportunist diseases, PLS, logit model, principal component analysis.

1. Introduction

Acquired immune deficiency syndrome (AIDS) follows from damage to the immune system caused by the human immunodeficiency virus (HIV). This damage leaves individuals prone to opportunistic infections and tumors.

Grouping countries according to the relationship between AIDS and opportunistic disease incidence may provide a useful framework for epidemiological studies.

The data we had available was the number of new cases per disease and country year by year.

To treat this data we decided on a double approach. First we tried to express the influence of AIDS on attacks of opportunistic diseases. Since there are several of these, we decided to find the linear combination of incidences of opportunistic diseases best related with AIDS incidence. Moreover the incubation time had to be taken into account. Thus we found the linear combination of incidences of opportunistic diseases with maximum correlation with lagged AIDS incidence. The lags we considered were 1, 2, 3 and 4 years. Such a study was made for the 13 countries of the EU for which we had sufficient data. Thus, for each lag, we had a two-entry table with a row per country and a column for each disease. To condense the information in such tables we carried out principal component analysis, considering the values of coefficients as variables. Moreover, the more relevant is the first principal component, the more clearly defined structure exists on the data. Hence we could use this as criteria for the choice of lag. It was found that a lag of three years was most relevant. Oliveira and Mexia (2004) found the most significant lag to be two years.

The second approach considered each disease separately. We used the same data, adjusting a logit model for its incidence. In this model we had a country effect and a time effect. Thus, for each disease, we obtained an adjusted country effect for every country in the study. We now had a second country-by-disease array, with the difference that we now also had a column for AIDS. The time span for AIDS was adjusted to take into consideration the three-year lag we had identified. To condense the information on this array we again applied principal component analysis.

We now had the results of both approaches condensed into two first principal components. Treating these as variables, a third and last principal component analysis was carried out, producing what we will call second-order principal components. We carried out this study in order to combine the results of the two approaches. As we shall see, this unification led to an interesting representation of the countries (Sequeira et al. 2008).

2. Methodology

As stated above, we used a two-fold approach to obtain the framework of the geometrical representation of the countries.

2.1. Retro-PLS

Firstly we found the linear combination of the incidences of a set of opportunistic diseases with maximum correlation with AIDS incidence. To take account of the AIDS incubation period we introduced lags of 1, 2, 3 and 4 years between the AIDS incidence and the opportunistic diseases incidences. Thus for each country and lag we had a vector of coefficients. Next, we used these vectors to obtain a geometric representation of the countries.

Let X be the number of registered AIDS cases per year and country, while $Y_1,\ldots,Y_k$ will be the numbers of registered cases, also per year and country, of opportunistic diseases. In our approach X will be a controlled variable and the others dependent variables. We have one controlled variable and several dependent variables, therefore we designated this method Retro-PLS.

When we work with lag l, the observations of X must precede by that number of years those of $\mathbf{Y}_k$ to give the components of the vectors

$$\mathbf{Z}_{k+1} = \begin{bmatrix} X & \mathbf{Y}_k^t \end{bmatrix}^t$$

Let us see how to carry out the first approach in our method.

Since the variance covariance matrix of vector $\mathbf{Y}$, $\Sigma(\mathbf{Y})$, is symmetric there will be an orthogonal matrix $\mathbf{P}$ that diagonalizes it, so that

$$\mathbf{P}\Sigma(\mathbf{Y})\mathbf{P}^t = D(v_1,\ldots,v_k)$$

where $D(v_1,\ldots,v_k)$ is the diagonal matrix whose principal elements are the eigenvalues of $\Sigma(\mathbf{Y})$. Moreover the line vectors $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_k$ of $\mathbf{P}$ will be the eigenvectors of $\Sigma(\mathbf{Y})$.

Taking $\mathbf{Y}^0 = \mathbf{PY}$ we get $\Sigma\left(\mathbf{Y}^0\right) = D(v_1,\ldots,v_k)$ and $\Sigma\left(\mathbf{Y}^0, X\right) = \mathbf{P}\Sigma(\mathbf{Y}, X)$ with $\Sigma(\mathbf{Y}, X)$ the cross covariance matrix of $\mathbf{Y}$ and X.

Thus, we can write $\mathbf{a}^0 = \mathbf{Pa}$ with $\mathbf{a}^0 = \Sigma\left(\mathbf{Y}^0, X\right)$ and $\mathbf{a} = \Sigma(\mathbf{Y}, X)$.

Since, when $\sigma^2(U) = 1$,

$$\operatorname{cov}(U,V) = corr(U,V)\sigma(U)\sigma(V) = corr(U,V)\sigma(V),$$

maximizing $corr(U,V)$ for a given V is equivalent to maximizing $\operatorname{cov}(U,V)$, since $\sigma(V)$ is fixed. Thus given $\mathbf{W} = \mathbf{w}^t\mathbf{Y}^0$ we intend maximizing

$$\Sigma\left(\mathbf{w}^t\mathbf{Y}^0, X\right) = \mathbf{w}^t\,\Sigma(\mathbf{Y}^0, X) = \mathbf{w}^t\mathbf{a}^0 = \mathbf{a}^{0^t}\mathbf{w} = \sum_{i=1}^{k} a_i^0 w_i$$

under the restriction

$$\boldsymbol{\Sigma}\left(\mathbf{w}^t\mathbf{Y}^0\right)=1 \Leftrightarrow \mathbf{w}^t\boldsymbol{\Sigma}\left(\mathbf{Y}^0\right)\mathbf{w}=1 \Leftrightarrow \mathbf{w}^t D(v_1,\dots,v_k)\mathbf{w}=1$$
$$\Leftrightarrow \sum_{i=1}^{k} v_i w_i^2 = 1 \Leftrightarrow \sum_{i=1}^{k} v_i w_i^2 - 1 = 0.$$

Using the method of Lagrange multipliers we obtain

$$w_i = \frac{a_i^0 v_1}{a_1 v_i} w_1 \text{ ; } i = 2,\dots,k \text{ with } w_1 = \pm \sqrt{\frac{1}{\sum_{i=1}^{k} \frac{\dot{a}_i^{02} v_1^2}{a_1^2 v_i}}} .$$

These are the coefficients of the linear combination of incidences of opportunistic diseases with maximum correlation with AIDS incidence for each country and lag. Then for each lag we consider a matrix $\mathbf{M}=[m_{i,j}]$.

The rows of this matrix will correspond to countries, and the columns to opportunistic diseases. Then $m_{i,j}$ will be the coefficient for the incidence of the *j*th disease in linear combination with the maximum correlation with AIDS for the *i*th country and that lag.

Then we work with the matrix $\mathbf{M}$ as with a data matrix, where the columns correspond to variables and the rows to objects. In our case the variables correspond to the opportunistic diseases and the objects to countries.

Next, we consider the principal components of these variables, obtaining the eigenvalues $\theta_1,\dots,\theta_k$ and the eigenvectors $\gamma_1,\dots,\gamma_k$ of the estimated covariance matrix of $\mathbf{M}$, $\sum(\mathbf{M})$.

We consider that the most relevant lag is the one in which $\sum(\mathbf{M})$ has the largest first eigenvalue. The columns of the matrix $\mathbf{B}=\mathbf{M}[\gamma_1 ... \gamma_k]$ contain the values of the principal components. The relative amounts of information carried by them may be measured by the ratios

$$r_h = \frac{\theta_h}{\sum_{i=1}^{k} \theta_i},\ h=1,\dots,k$$

which enable us to decide how many principal components are relevant.

Moreover the rows of $\mathbf{B}$ correspond to countries, so we may use the values of the most relevant principal components to obtain geometrical representations of the countries.

2.2. Logit model

Secondly, for the incidences of each disease, AIDS and all associated diseases, logit models were adjusted. In these models we consider two factors, country and year, assuming they were additive.

We thus assumed that

$$y_{i,j} = \text{logit}(p_{i,j}) = \ln\frac{p_{i,j}}{1-p_{i,j}} = \alpha + \beta\left(f_i + g_j\right)$$

with $p_{i,j}$ the probability of an individual's being infected with some disease in country i during year j, the coefficients f_i corresponding to countries, $i = 1,\ldots,m$ and the g_j, $j = 1,\ldots,n$, to years.

When we work with lag l, which in our case is that determined in the previous section as most relevant, the incidence of AIDS must precede the incidence of the associated diseases by that number of years.

We now obtain a matrix $\mathbf{F} = [f_{i,k}]$ with a row for each country and a column for each disease. The element in the *i*th row and *k*th column will be the country effect $\left(f_i\right)$ for the *i*th country and *k*th disease. Thus we are considering variables associated to diseases. We now apply a principal component analysis to the data in **F**.

2.3. Second-order PCA

Finally, we considered the first component derived from the retro-PLS and from logit models as new variables. We standardized them and applied a principal component analysis. Both new principal components are then used to obtain a final geometrical representation of the countries.

3. Application

3.1. Data

We used the data available in WHO/Europe's statistical databases: Centralized Information System for Infectious Diseases (CISID) and European Health For All Database (HFA-DB). We worked with the number of notified AIDS cases and the number of new cases of five associated diseases $(k = 5)$, per year, in the first 15 member countries of the European Union.

The number of years varied from country to country, according to the available data, as we can see in Table 1.

Table 1. Number of years per country

Country	Years
Austria	1992 to 2003
Belgium	1987 to 2000
Denmark	1982 to 2003
Finland	1989 to 2003
France	insufficient data on Hepatitis
Germany	1987 to 2003
Greece	1982 to 2001
Ireland	1982 to 2001
Italy	1983 to 2002
Luxembourg	1989 to 2004
Netherlands	insufficient data on Salmonellas
Portugal	1990 to 2003
Spain	1993 to 2004
Sweden	1989 to 2003
United Kingdom	1984 to 2003

Due to lack of data it was not possible to include either France or the Netherlands. Again on the basis of availability of data, we selected the following associated diseases: Tuberculosis, Salmonella, Hepatitis A, Hepatitis B and Hepatitis originated by other viruses.

We applied the described methodology in the previous section, considering a lag of 1, 2, 3 or 4 years.

3.2. Results

In Table 2 we present the relative amounts of the information carried by the first two principal components from Retro-PLS, for different lags.

Table 2. Relative information of the principal components

	r_1	r_2
lag 1	0.840	0.100
lag 2	0.734	0.153
lag 3	0.959	0.028
lag 4	0.798	0.094

We can observe that with a lag of 3 years, the first eigenvalue represents 96% of the information and the second 3%, so this will be the most relevant lag.

The adjusted coefficients for lag 3 are displayed in Table 3.

Table 3. Adjusted coefficients for lag 3

	Tuberculosis	Hepatitis A	Hepatitis B	Viral Hepatitis	Salmonellosis
Austria	0,000110899	-0,004102972	-5,90324E-05	0,002429236	0,012036322
Belgium	7,42317E-05	-0,00395183	-0,003071526	0,002099724	-0,001225302
Denmark	0,000984927	0,002094747	0,004160311	0,001996282	-0,005700696
Finland	9,17767E-05	-0,00061138	0,000311577	0,003258527	-0,004717394
Germany	1,40374E-05	0,00030253	-0,000323601	-0,000574693	-0,000469933
Greece	0,001326813	-0,000869046	-0,003273839	-0,002372444	-0,002485704
Ireland	0,001941656	0,000320938	-0,006732749	0,003244112	-0,002334454
Italy	3,20302E-05	8,90536E-05	-0,000414864	0,000538815	-0,000324801
Luxembourg	0,003820423	0,004077511	-0,006882235	0,015209122	-0,085468448
Portugal	0,000792664	0,001785793	0,000764782	-0,002895699	0,00177981
Spain	0,000392095	-0,000159805	-4,03259E-05	1,38491E-05	0,002771008
Sweden	0,000626509	0,000735095	0,004371187	-0,00915917	-0,006769841
UK	2,98069E-05	0,000193965	-8,48745E-05	-0,000458494	-0,004263992

In Figure 1 we show the geometrical representation given by the two first principal components. Observe that Luxembourg appears isolated.

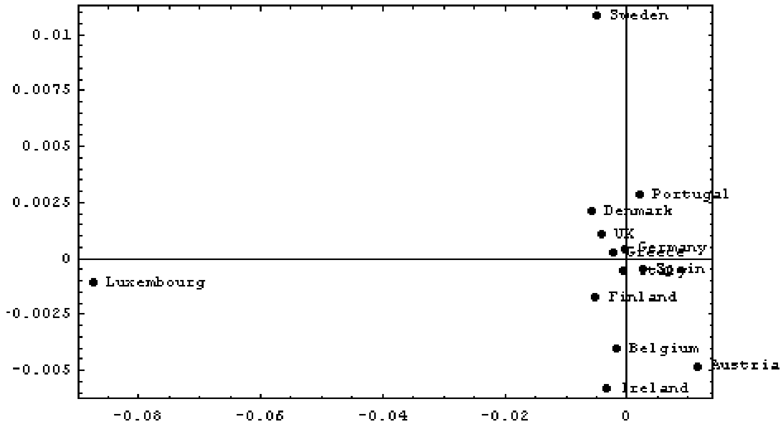

Figure 1. First and second retro-PLS principal components

Concerning the results from logit model: In the logit models, as stated above, a 3-year lag was used.

Now we will present the results from a logit model taking the incidences of AIDS three years before those of the opportunistic diseases. We analyzed the AIDS incidences over six years $(n=6)$, from 1996 to 2001. We applied the zigzag algorithm to obtain the adjusted coefficients.

In Table 4 we present the adjusted coefficients f_i; $i=1,\ldots,13$ for each disease.

Table 4. Local factors adjusted coefficients

	AIDS	Tuberculosis	Hepatitis A	Hepatitis B	Viral Hepatitis	Salmonellosis
Austria	-3,85237	-2,70924	-3,69846	-2,97233	-2,48524	-1,55890
Belgium	-3,7102	-2,77077	-3,29774	-2,87946	-2,86580	-1,26371
Denmark	-3,49317	-3,06730	-3,74857	-4,37421	-2,68246	-2,26864
Finland	-5,0901	-3,04974	-3,18430	-3,40679	-1,22775	-2,12899
Germany	-3,85371	-3,00417	-3,66479	-3,37315	-1,86336	-1,56854
Greece	-3,73799	-3,42660	-3,91962	-3,82653	-3,98228	-3,93032
Ireland	-4,271	-2,91352	-3,08505	-2,48924	-2,70270	-3,35664
Italy	-2,47707	-3,30738	-3,47378	-3,76742	-3,59268	-3,09743
Luxembourg	-3,31991	-3,07592	-2,16203	-2,44186	-1,06132	-1,57825
Portugal	-1,77119	-1,53043	-4,22090	-3,73089	-3,44098	-4,33749
Spain	-1,78909	-2,33674	-3,76304	-3,77820	-3,19482	-3,11203
Sweden	-4,19897	-3,68993	-4,06741	-3,40167	-1,09136	-2,16378
UK	-3,57133	-2,87139	-3,81861	-4,04434	-2,57961	-2,79531

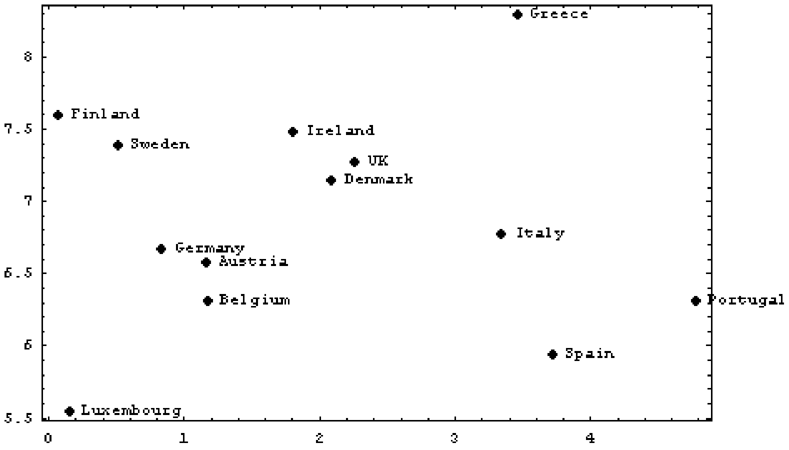

Figure 2. First and second logit models principal components

We can observe in Figure 2 the first and second principal components of local factors. In this case the first eigenvalue represents 60.5% of the information and the second 15%.

Lastly, in Figure 3, we show the geometrical representation of the countries using the second-order principal components.

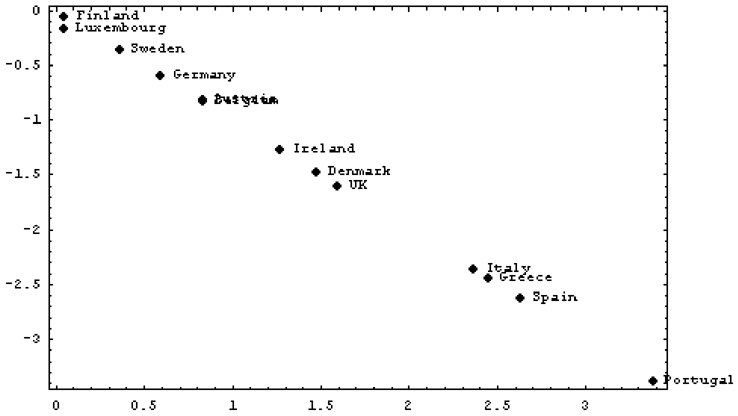

Figure 3. Second-order principal components

3.3. A numerical experiment

Since Luxembourg appears isolated in Figures 1 and 2 and it has the smallest population, it would be interesting to test what would happen if we increased its population 50 and 100 times, keeping the same rates of infection.

The results of this experiment are presented in Figure 4.

Observe that in both cases the fictitious country appears very near the real Luxembourg and the grouping of the countries is maintained.

To complete our treatment we computed the WALD statistics for the hypothesis of no significant difference between Luxembourg and 50×Luxembourg and Luxembourg and 100×Luxembourg.

The values of these statistics were 1.02 and 1.16. None of them are significant. Moreover these statistics would, under the null hypothesis, be central chi-squares with 1 degree of freedom.

Thus the obtained values for the statistics are very near their mean value (1) under the null hypothesis.

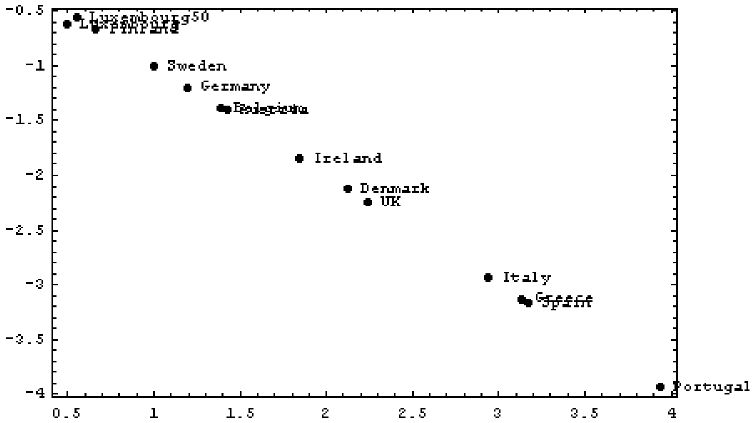

Figure 4. Second-order principal components with 50×Luxembourg

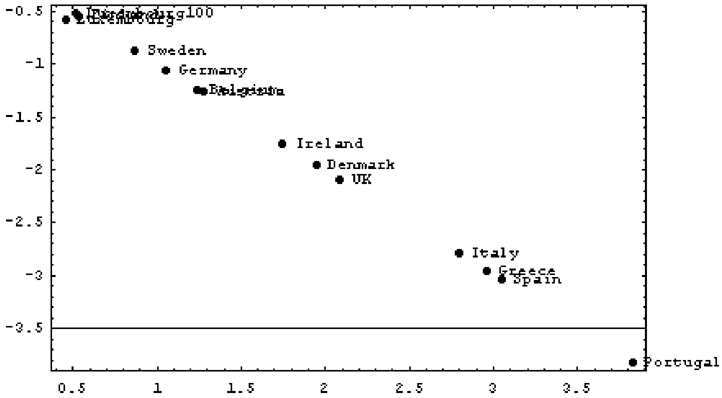

Figure 5. Second-order principal components with 100×Luxembourg

4. Comments and conclusions

Using the technique described above we found that the most representative lag between AIDS and the five opportunistic diseases was 3 years.

The countries could now be grouped into:

1. Portugal, isolated;

2. The group nearest to Portugal, consisting of southern European countries: Italy, Greece and Spain;
3. The remaining countries, grouped in a third cluster.

As stated in the introduction, the grouping of the countries may lead to further studies. These would attempt, for each group of countries, to find the relevant factors behind the dynamics of opportunistic AIDS-related diseases.

This grouping was validated by an experiment in which we increased the population of Luxembourg 50 and 100 times while keeping the same incidence rates.

Acknowledgement

The authors wish to express their sincere thanks to Professor Christoph Minder of Berne University for his suggestion concerning Luxembourg.

REFERENCE

Jolliffe I.T. (2002): Principal Component Analysis, Springer, 2nd edition.

Oliveira, M.M., Mexia J.T. (2004): AIDS in Portugal: endemic versus epidemic forecasting scenarios for mortality, International Journal of Forecasting 20: 131-135.

Sequeira I.J., Mexia J.T., Nunes S. (2008): Double Minimization for Logit Models with an Additive Two Factors Structure, Biometrical Letters 45(1): 69-80.

WHO/Europe - Data, Statistical data: European Health For All Database (HFA-DB): http://www.euro.who.int/hfadb

WHO/Europe - Data, Statistical data: Centralized Information System for Infectious Diseases (CISID): http://data.euro.who.int/cisid