

R: A valuable tool for linking high education and research.

GABRIEL IBARRA-BERASTEGI^{1,2}, UNAI GANZEDO^{1,3}, JON SAENZ^{1,3},
AGUSTIN EZCURRA^{1,3}, IÑIGO ERRASTI^{1,2}, ANA ELIAS^{1,4}, ASTRID
BARONA^{1,4}, LEYRE INSAUSTI^{1,4}

⁽¹⁾ University of the Basque Country

⁽²⁾ Dept. of Fluid Mechanics

⁽³⁾ Dept. of Applied Physics

⁽⁴⁾ Dept. of Environmental and Chemical Engineering

Bilbao's Engineering School

Alda. Urkijo s/n 48013. Bilbao.

SPAIN

gabriel.ibarra@ehu.es <http://www.ehu.es/eolo>

Abstract:- The use of non-commercial software is sometimes disregarded and not properly considered in all its implications and possibilities for scientific research and higher education. One example is R, is a freely available software which is increasingly being used in a wide range of scientific fields and adopted by many institutions as a standard. It can be a most valuable tool to bring together the efforts from all the scientific community, ranging from students to senior scientists worldwide. Two examples corresponding to two research groups from the University of the Basque Country (Spain) are shown. The conclusion is that R is an excellent tool to bridge high education and research and may represent a valuable contribution to the advantage of scientific knowledge.

Key-Words: - high education, research, R, fluid mechanics, applied physics, chemical engineering

1 Introduction

The use of non-commercial software is sometimes disregarded and not properly considered in all its implications and possibilities for scientific research and higher education. This is particularly true when we come to the wide range of educational possibilities that freely available software may represent. One example of this is R, a freely available software <http://www.r-project.org/> which is increasingly being used in a wide range of scientific fields and adopted by many institutions as a "standard". R is a high-level language like Matlab, Mathematica or IDL intended for generic data analysis tasks common in scientific work and publishing-quality graphical representation.

R uses a basic module (currently version R 2.9.1) with a set of basic statistical applications and features. Additionally, specific packages which have been developed by research groups from several scientific fields can be downloaded <http://lib.stat.cmu.edu/R/CRAN/> since they are offered freely to the rest of the scientific community.

This includes a great deal of scientific tools used in simulation and data mining in a number of fields such as engineering, meteorology, genetics, climate, agriculture, biology, environmental protection, linguistics to topography, to mention a few. The tools include genetic algorithms, machine learning,

multiple linear regression, neural networks, self-organizing maps, cluster analysis, principal component analysis. R is an object-oriented software which runs using scripts that are quite simple to write. Besides, examples are easily available in the web for a huge number of applications. In a worldwide mentioned article in the New York Times (7-Jan-09) this is what was written on R :

"R is the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

But R has also quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use. R is really important to the point that its hard to overvalue it, said Daryl Pregibon, a research scientist at Google, which uses the software widely. It allows statisticians

to do very intricate and complicated analyses without knowing the blood and guts of computing systems.

It is also free. R is an open-source program, and its popularity reflects a shift in the type of software used inside corporations. Open-source software is free for

anyone to use and modify. I.B.M., Hewlett-Packard and Dell make billions of dollars a year selling servers that run the open-source Linux operating system, which competes with Windows from Microsoft. Most Web sites are displayed using an open-source application called Apache, and companies increasingly rely on the open-source MySQL database to store their critical information. Many people view the end results of all this technology via the Firefox Web browser, also open-source software.

R is similar to other programming languages, like C, Java and Perl, in that it helps people perform a wide variety of computing tasks by giving them access to various commands. For statisticians, however, R is particularly useful because it contains a number of built-in mechanisms for organizing data, running calculations on the information and creating graphical representations of data sets.

Some people familiar with R describe it as a supercharged version of Microsofts Excel spreadsheet software that can help illuminate data trends more clearly than is possible

What makes R so useful and helps explain its quick acceptance is that statisticians, engineers and scientists can improve the softwares code or write variations for specific tasks. Packages written for R add advanced algorithms, colored and textured graphs and mining techniques to dig deeper into databases.

Close to 1,600 different packages reside on just one of the many Web sites devoted to R, and the number of packages has grown exponentially. One package, called BiodiversityR, offers a graphical interface aimed at making calculations of environmental trends easier.

Another package, called Emu, analyzes speech patterns, while GenABEL is used to study the human genome.

The financial services community has demonstrated a particular affinity for R; dozens of packages exist for derivatives analysis alone.

The great beauty of R is that you can modify it to do all sorts of things, said Hal Varian, chief economist at Google. And you have a lot of prepackaged stuff thats already available, so youre standing on the shoulders of giants.

R first appeared in 1996, when the statistics professors Ross Ihaka and Robert Gentleman of the University of Auckland in New Zealand released the code as a free software package.

According to them, the notion of devising something like R sprang up during a hallway conversation. They both wanted technology better suited for their

statistics students, who needed to analyze data and produce graphical models of the information. Most comparable software had been designed by computer scientists and proved hard to use.

Lacking deep computer science training, the professors considered their coding efforts more of an academic game than anything else. Nonetheless, starting in about 1991, they worked on R full time. We were pretty much inseparable for five or six years, Mr. Gentleman said. One person would do the typing and one person would do the thinking.

Some statisticians who took an early look at the software considered it rough around the edges. But despite its shortcomings, R immediately gained a following with people who saw the possibilities in customizing the free software.

John M. Chambers, a former Bell Labs researcher who is now a consulting professor of statistics at Stanford University, was an early champion. At Bell Labs, Mr. Chambers had helped develop S, another statistics software project, which was meant to give researchers of all stripes an accessible data analysis tool. It was, however, not an open-source project.

The software failed to generate broad interest and ultimately the rights to S ended up in the hands of Tibco Software. Now R is surpassing what Mr. Chambers had imagined possible with S.

The diversity and excitement around what all of these people are doing is great, Mr. Chambers said.

R has really become the second language for people coming out of grad school now, and theres an amazing amount of code being written for it, said Max Kuhn, associate director of nonclinical statistics at Pfizer. You can look on the SAS message boards and see there is a proportional downturn in traffic.

SAS says it has noticed R's rising popularity at universities, despite educational discounts on its own software, but it dismisses the technology as being of interest to a limited set of people working on very hard tasks.

I think it addresses a niche market for high-end data analysts that want free, readily available code, said Anne H. Milley, director of technology product marketing at SAS.

She adds, We have customers who build engines for aircraft. I am happy they are not using freeware when I get on a jet. While it is difficult to calculate exactly how many people use R, those most familiar with the software estimate that close to 250,000 people work with it regularly. The popularity of R at universities could threaten SAS Institute, the privately held business software company that specializes in data analysis software. SAS, with more than \$2 billion in annual revenue, has been the preferred tool of scholars and corporate managers.

But while SAS plays down Rs corporate appeal, companies like Google and Pfizer say they use the software for just about anything they can. Google, for example, taps R for help understanding trends in ad pricing and for illuminating patterns in the search data it collects. Pfizer has created customized packages for R to let its scientists manipulate their own data during nonclinical drug studies rather than send the information off to a statistician.

The co-creators of R express satisfaction that such companies profit from the fruits of their labor and that of hundreds of volunteers.

Mr. Ihaka continues to teach statistics at the University of Auckland and wants to create more advanced software. Mr. Gentleman is applying R-based software, called Bioconductor, in work he is doing on computational biology at the Fred Hutchinson Cancer Research Center in Seattle.

R is a real demonstration of the power of collaboration, and I dont think you could construct something like this any other way, Mr. Ihaka said. We could have chosen to be commercial, and we would have sold five copies of the software”.

Here is how MJ Venables one of its developers, belonging to the so called R development group provides the following description on what R is:

«R is an integrated suite of software facilities for data manipulation, calculation and graphical display. Among other things it has

- *an effective data handling and storage facility,*
- *a suite of operators for calculations on arrays, in particular matrices,*
- *a large, coherent, integrated collection of intermediate tools for data analysis,*
- *graphical facilities for data analysis and display either directly at the computer or on hardcopy,*
- *and*

a well developed, simple and effective programming language (called ‘S’) which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.)The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R are essentially

ephemeral, written for a single piece of data analysis. [1]»

One of the most important features regarding R is the fact that is the result of a collaborative effort made by the scientific community as such. Futhermore, any user could write his/her own package. No matter how specific these applications may be, most probably they can be useful to someone else.

The question may arise as to why switch to using R when you have mastered a perfectly adequate statistical package already?. It is commonly accepted that the main advantage of R [2] is that *«any member of the scientific community (and that include also students!) can take advantage of its unrivalled coverage and the availability of cutting edge applications»*. This is particularly true in the case of very recent applications like those in the field of machine learning (Relevance Vector Machine, Random Forests and others).

A second powerful reason is that R is being adopted as a standard by an increasing number of research groups and institutions. This can be seen in the scientific literature where an increasing number of papers are the result of calculations and graphs carried out in R. Adoption of R may contribute to “speak the same language” that your colleagues worldwide, thus enabling a much easier exchange of results. For example, in the case of fisheries management, the scientific community has developed the FLR library. The FLR library is a collection of tools in the R statistical language <http://r-project.org> that facilitates the construction of bio-economic simulation models of fisheries and ecological systems. It is a generic toolbox, but is specifically suited for the construction of simulation models for the evaluation of fisheries management strategies. The FLR library is under development by researchers across a number of laboratories and universities in various countries. <http://flr-project.org>. Similar packages can be found for a great number of specific scientific fields at the CRAN repository <http://lib.stat.cmu.edu/R/CRAN/>.

Last but not least, R is free though some commercial versions exist. In these versions particular aspects of R are improved, mainly those intended to provide a more friendly environment for the final use of the tool in particular fields [3][4][5][6].

Every year there is a conference of R users <http://www2.agrocampus-ouest.fr/math/useR-2009/> (Fig. 1) in which advances in the use of R are shown and incorporated into the main repository of knowledge. In an increasing number of very specific fields, R is being adopted as a standard to make easier the exchange of information and advance of knowledge.



Fig. 1. Logo of the R users conferences

An important aspect when used in Universities, is that the use of R blurs the differences between high-level students and researchers, since R can be used simultaneously for educational and research purposes. Since programming in R is done using scripts that invoke certain functions belonging to the basic module and/or to the packages, the work can be iteratively improved at the different stages. This makes R the perfect tool to incorporate the efforts of students and researches in an iterative way to the common research effort.

The purpose of this paper is to communicate the experience of two research groups at the University of the Basque Country (Spain) regarding the adoption of R –partially or fully- as a standard in their research projects under development.

These two research groups work in different environmental issues. One of them is the Eolo Group, formed by different senior scientist, researchers, students and professionals, all of them holding links with the University. The main working areas of Eolo could be defined as “Meteorology, Climate and Environment” with several papers and research projects currently going on. <http://www.ehu.es/eolo>, particularly including the diagnosis of the performance of global climate models around the Iberian Peninsula. One of the major current projects of Eolo is the estimation of future climate scenarios for the Basque Country (Spain).

The second group is the biofiltration group with a similar structure and whose main research activities focus on the use of microorganisms to eliminate pollutants from a gaseous flow.

As a first example, the use of R in a research project currently being carried out by Eolo to analyse the performance of global climate models in North Atlantic and the Iberian Peninsula is shown. The second example, is the calculation of the slopes that represent the degradation speed of pollutants by microorganisms. Comparison of slopes from different bioreactors under different operating conditions

allows the identification of optimal conditions leading to the best degradation rates.

2 Problem Formulation

As mentioned above one of the major projects currently going on is the diagnosis of the performance of global climate models in North Atlantic and the Iberian Peninsula.

The main activities in this project could be divided into two groups:

a) Very specific short courses directly related to the concepts needed to carry out the main objectives in the research project. To that end high-level and internationally recognized researchers have been brought to the University of the Basque Country. This has also made it possible to strengthen ties with their institutions worldwide. The total amount of courses offered has been up to now 6.

b) The specific tasks directly aiming to obtaining the results of the project, amongst others, to diagnose the performance of the global climate models in simulating past climate around the Iberian Peninsula.

In both cases, R has been adopted as a standard, thus making it possible the exchange of information inside the group and also with other groups worldwide. It is important to mention that important tasks needed to obtain the final objectives have been carried out using specific R packages like the analogues and *clim.pact*, freely available at <http://lib.stat.cmu.edu/R/CRAN/>. R scripts needed to carry out calculations have been obtained as the fruit of a collaborative effort among students –previously familiar with R during their doctorate studies- and senior researches. Also, important parts of the scripts have been obtained from previous works which after modifications have been easily incorporated to the final scripts which, run in the frame of R have made possible to reach the final results.

The first example is the identification of the AR4 models that best represented surface level pressure in the 20th century. A total amount of 69 runs of different models from the AR4 repository have to be evaluated by comparing how properly have they been able to describe observed SLP during the 20th century. For different areas (Fig. 2a-2b) this performance can be estimated to represent the model's ability (Table 1) that can be expected in the simulations of climate during the 21st century.

This may be helpful to obtain feasible scenarios and therefore predictions yielded by those finally selected models, which can inspire future mitigation initiatives against the effects of climate change. In the case of the biofiltration group, several bioreactors (Fig. 3)

have been put up operating under different conditions.

The use of R allows graphical representation of raw data obtained from the different biofilters. At a second step, slopes or degradation rates can be calculated.

INSTITUTION	AREA
NOAA. USA	<i>1</i>
CCSR. JAPAN	<i>2</i>
CCSR. JAPAN	<i>3</i>
NOAA. USA	<i>4</i>
NOAA. USA	<i>5</i>
CCSR. JAPAN	<i>6</i>
CCSR. JAPAN	<i>7</i>
NASA. USA	<i>8</i>
MRI. JAPAN	<i>9</i>
Max Planck Institute. GERMANY	<i>10</i>
MRI. JAPAN	<i>11</i>
NOAA. USA	<i>12</i>
CCSR. JAPAN	<i>13</i>

Table 1. Best model (20C3M) for the different areas of Fig. 2a-2b.

It is obvious for both cases that there is a continuity between high education, research and inner short-courses in the frame of the research groups that can be attributed to the adoption of R –fully or partially– as a standard.

The results of this decision are highly positive and provide a great number of advantages regarding the everyday operative exchange of scientific results, educational efforts and initiatives inside the groups.

3 Problem Solution

For the first example, the code that must be written is mainly the one related to the input-output of data (see Appendix), but strictly, the statistical part is straightforward.

The calculations are not difficult and involve the estimation of correlation coefficients as well as the comparison of probability density functions using the Kolmogorov-Smirnov test.

In R, this is very simple and can be done with one single command

For correlation coefficient, using the cor function

```
res<-cor(slpfileseamod[,j],slpfileseaobs[,j])
```

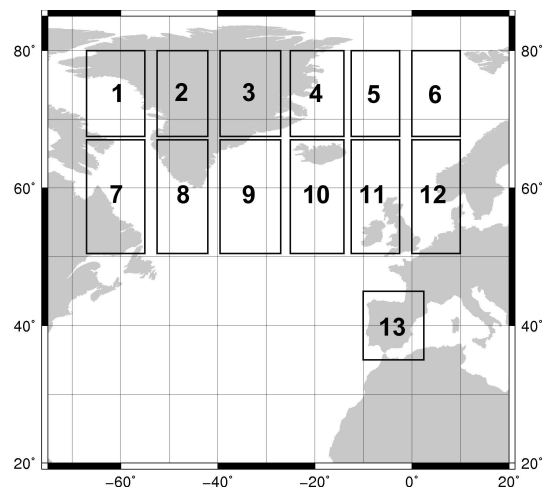


Fig.2a. Areas for which the 69 AR4 models have been evaluated

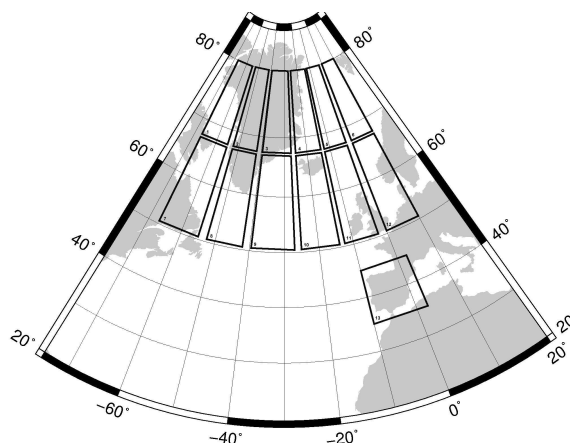


Fig.2b. The same as Fig. 2a in Lambert coordinates

For the Kolmogorov-Smirnov test with the **ks.test** function

```
slpres<-ks.test(slpfileanomod,slpfileanoobs)$p.value
```

The best models are those:

- i) with a higher correlation coefficient between observed and modelled sea-level pressure seasonal cycles
- ii) whose modelled anomalies are not different at a 95% confidence level from those observed in the 20th century

The final ranking can be seen in table 1

The second example involves:

- i) Graphical representation of raw data from different experiments correspondings to biofilters operating under different conditions (biofilters #5, #6, #7, #8, Fig. 4a-4d)
- ii) Calculation of slopes, that is, degradation speed of pollutant by microorganisms under different operating conditions
- iii) Intercomparison of slopes in order to determine at a 95% confidence level whether the apparently different values of the slopes are truly different or not

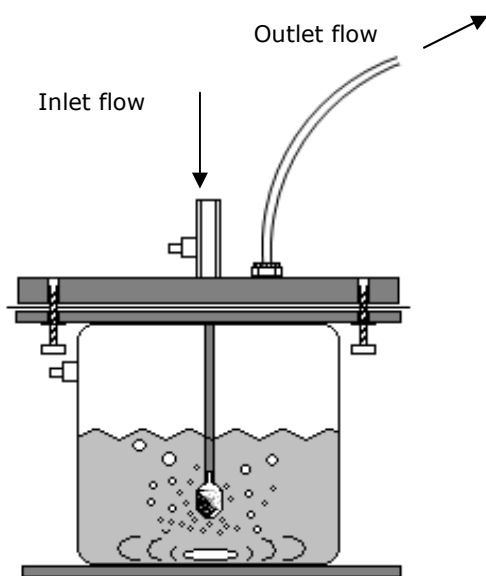


Fig. 3. Diagram of the biofiltration systems (continuous acclimation).

For the graphs, two commands are enough

```
plot(x1,y1,xlab="Degradation time (min)",ylab="Concentration",main="Reactor N#1")
```

to obtain the representation of data and

```
abline(lm(y1~x1))
```

to graph the regression line (Fig. 4a-4d)

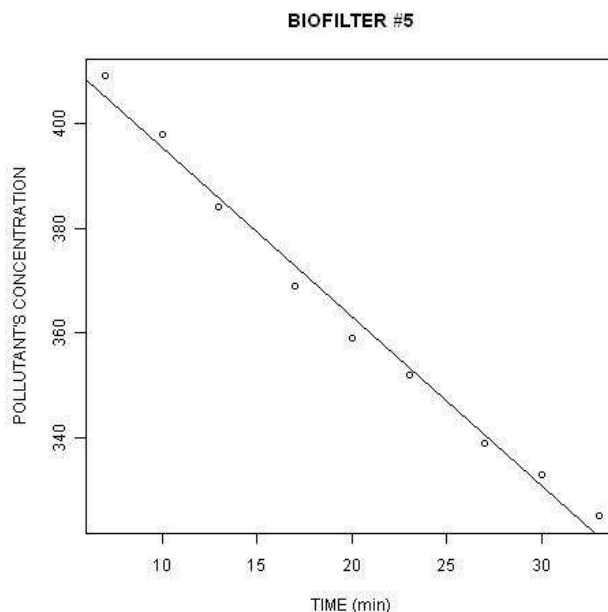


Fig. 4a. Graphical representation of the contaminant degradation due to the action of microorganisms in reactor #5.

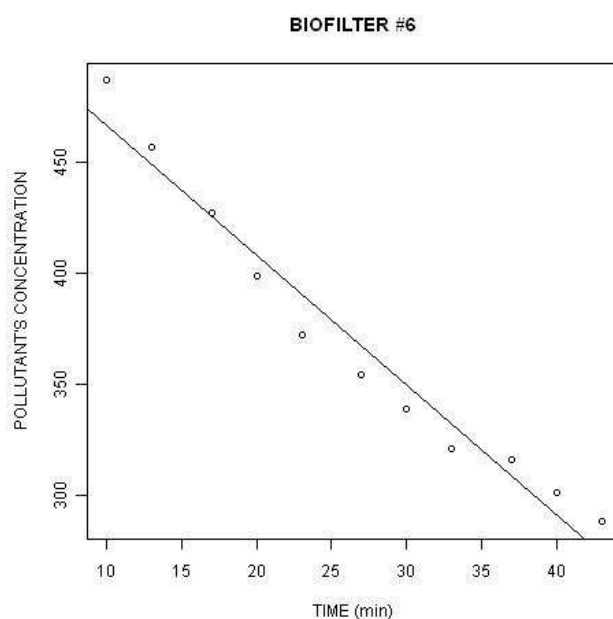


Fig. 4b. Graphical representation of the contaminant degradation due to the action of microorganisms in reactor #6.

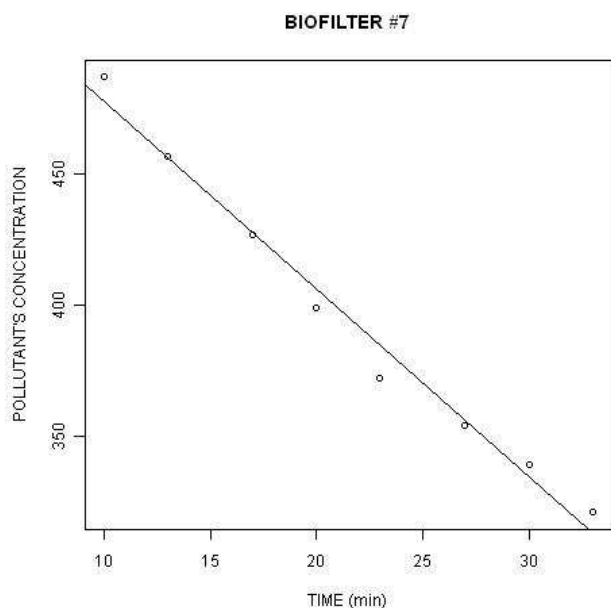


Fig. 4c. Graphical representation of the contaminant degradation due to the action of microorganisms in reactor #7.

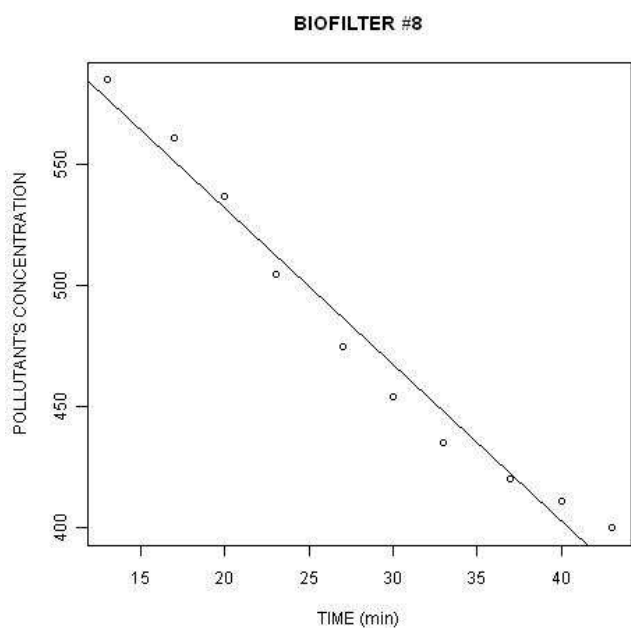


Fig. 4d. Graphical representation of the contaminant degradation due to the action of microorganisms in reactor #8.

Finally, to obtain the slope of the regression

```
modelreactor5<-lm(conc5~deptime5)
slopereactor5<-coef(modelreactor5)[2]
```

In case we would also like to calculate the intercept of the regression line, this can be easily done in R by changing the value of the coefficient from [2] to [1] and writing the following line for the object interceptreactor5

```
interceptreactor5<-coef(modelreactor5)[1]
```

	Slope (degradation rate)
Reactor #5	-3.22
Reactor #6	-5.85
Reactor #7	-7.16
Reactor #8	-6.43

Table 2. Degradation rates (slopes) in the biofilters #5, #6, #7, #8.

The third step involves solving the problem of assessing whether the average values of the slopes (shown in Table 2) are different at a 95% confidence level.

This is important because the different slopes are associated to different operating conditions and the whole experiment with different reactors is intended to identify under which conditions the microorganisms degrade higher fractions of pollutants.

In the cases where average values of slopes are not very different it is necessary to assess whether they are truly different with a certain degree of certainty, usually 95%.

Due to the low number of cases used (typically 10-15), the plain comparison of the slopes obtained might be highly misleading. This applies to the intercomparison among slopes corresponding to the comparison of slopes among different operation conditions.

If the slopes corresponding to different essays is to be compared, the plain use of the values calculated can be helpful for a preliminary assessment. However, if the values of these slopes obtained for two essays are quite similar, 95% confidence levels should be

