Enhancing Statistics Teaching with a Virtual Lab

A Case Study of Seamless Local and Remote Computing

Araceli Garín¹, Fernando Tusell*², Aitziber Unzueta³

^{1,*2}Facultad de CC.EE. y Empresariales, University of the Basque Country (UPH/EHU) Avenida Lehendakari Aguirre, 83, 48015 Bilbao, Spain

³E.U. Ingeniería Técnica Industrial, University of the Basque Country (UPH/EHU) Paseo Rafael Moreno "Pitxitxi", 2, 48013 Bilbao, Spain

¹mariaaraceli.garin@ehu.es; ^{*2}fernando.tusell@ehu.es; ³aitziber.unzueta@ehu.es

Abstract

Statistics and Data Analysis courses demand a fair amount of computing. While there are excellent free source tools which can be given away to students, their seamless integration requires a fair amount of work and is challenging to the less computer-savvy students. In the past we have addressed the problem by compiling and integrating the necessary tools in CD-ROM's and providing local computing facilities, but this has proved impractical on a number of counts. Trial and error has led us to finding a setup with which we have finally solved most problems.

Keuwords

E-learning; Remote Computing; Virtual Lab; Client-server

Introduction

Teaching Statistics and related subjects such as Econometrics or Operations Research to students of Economics presents various challenges; not least, applied courses on these subjects require a fair amount of practice, all of which entails the use of a computer.

We see it as an integral part of our duties to help students gather practical experience in our subjects; practice brings about in students greater appreciation of the power and usefulness of Statistics. To this end, we encourage student's work with data and try to involve them in answering interesting, moderately challenging questions. Benefits include an improved, more alert attitude in the classroom, and increased opportunities to work cooperatively and furnish mutual help, when learning new computing skills or the use of new software tools.

Our subject, Statistics, is somewhat special for the reasons explained below. Statistics is Mathematics with an open window on real-world phenomena. Because it is Mathematics, it is ideally suited to computer experimentation which is not tied to a

physical location nor does it need measurement instrument's manipulation. And yet it is a subject that benefits from, and indeed *requires*, experimentation. This dual nature of the subject, at once mathematical and "real world", is what it makes it particularly suited for a virtual lab.

What we call "virtual lab" is not without precedents; in fact, at the time of writing (2013), a Google search with the terms "virtual labs educational aspects" returns close to 12 million hits. It is therefore interesting to compare some other experiences to ours and try to ascertain when, how, and why, a virtual lab can add substantially to the learning process.

In the old, pre-1990 days, we would provide access to a computer lab, typically made of a mainframe or minicomputer with terminals connected in its close vicinity. Any computer work would be done there or, rarely, from a distant location through a dial up line.

The advent and rapid dissemination of personal computers changed all this. For over a decade, the "PC room", a bunch of PC's possibly connected in a local area network, was the dominant paradigm in ours and many other schools.

In this paper we present an experience whose goal is to provide a seamless environment for statistical work both in local facilities and through remote connections. We also describe the particular combination of factors that have made our experience successful. Unlike in most natural sciences, statistical work can be done remotely just as well as locally, if students have general access to computer hardware and wide band Internet access. The nature of the subject, the general availability of hardware and Internet access and the existence of very high quality open source (or otherwise freely distributable) software, are the factors that together explain the success of our experience.

The remainder of the paper is organized as follows. In Section 2 we list the advantages and disadvantages of both the old computer lab model and the PC roombased instruction, as we see them. In Section 3 we describe our experience on how a dual-faceted computer Lab, fitted with PC's yet having a workstation as its cornerstone, provides the best of both models, increases the productivity of students and teachers, and can be run with a (relatively) modest input of skilled labour. We comment briefly on some details of the administration, the benefits to our students' learning, and discuss some other experiences in the use of virtual labs in Statistics or more generally in education. Section 4 presents some ideas on possible future developments, and Section 5 concludes.

The past

The Old-style Computer Room

The pre-1990 computer room, which was prevalent at universities all over the world until the final eighties of the last century, was not without advantages. The requirements of skilled labour were moderate: there was a single machine, or only a few, to be run. Dumb terminals, even X-terminals with graphical capabilities, were an "install-and-forget" type of task, and run until physically worn out.

There was no question of viruses spreading, no questions of illegal copying, and much reduced concerns about security: there was little an inexperienced user could do to cause damage. Security of information was also easier to enforce. In fact, a computer room with dumb terminals or "thin-clients" may be the choice even nowadays when security of the data is paramount: an interesting experience is the Virtual Microdata Lab of the UK Office for National Statistics (ONS), a description of which is given in Ritchie (2008).

Since dumb terminals were of little use to most users, hardware theft was also a rarity.

Finally, from the teaching standpoint which is what interests us here mostly, the physical proximity of students having the same problems afforded many opportunities for cooperative learning and mutual help.

The shortcomings, however, are also evident: essentially all work had to be done in one place, and users depended on computer staff even for the simplest tasks, such as bringing in new data whose size meant that it could not be typed directly at a

terminal.

The PC Room

The PC rooms changed all this. Suddenly, the users were masters of themselves. They found at the school the same hardware, the same operating system and software tools that they were used to work with at home. There was no retraining, no learning curve, no dependence on any one to move data, which could travel in floppies or, later, pen drives.

Alas, this very flexibility was not without inconvenients. PC's could be infected and their hard drives erased. In order to have shared services such as storage or printing, local area networks (LAN's) had to be set up. Linking different machines into the same LAN, managing authentication in a centralized manner, ensuring consistency and integrity of the software installed, protecting users from themselves and each other, became tasks that dwarfed the effort previously required for the administration of a single machine.

Since, in addition, hardware and software have short useful lives because of technical obsolescence, managing computer rooms is a task extremely demanding of resources. When all is taken into consideration, the advantage of standard, off-the shelf hardware is negated by the complexity of the installation and maintenance.

The "Help Yourself" Approach

The increase in availability and quality of free software opened new perspectives: software could be given away to students in a CD or DVD and, at least for work requiring only moderate resources, they were able to work at home or wherever they could bring their laptops. It seemed for a while that providing computer resources to students was no longer a problem.

It soon became evident that this was not the case. For one thing, preparing, customizing and integrating all the necessary tools is a lot of work —and work that needs to be redone frequently, so that the software stays reasonably current. On the other hand, installation in a variety of hardware, with different operating systems or different versions of the same operating system, requires much testing, is very errorprone and in our case meant that a sizeable proportion of the students failed to have working installations.

Another drawback is that commonly used and voluminous information (like digital cartography or

time series data collections) is difficult to share and keep current by way of handing over CD-ROM's.

From the point of view of learning, this approach also meant individual work, with greatly diminished opportunities for interaction among students—a factor whose importance cannot be overemphasized, as in our experience they learn much more easily applied skills by interacting among themselves than in isolation.

The Virtual Lab

Hardware and Software Setup

Around 1999 it became obvious that none of the approaches we had tried had been fully, or even moderately, successful. In an attempt to regain the advantages of the old-style computer lab and keep the advantages of modern PC's, we decided to mix both; the success has been above our expectations, and we think this success is the outcome of a delicate interaction of several factors, which we did not quite foresee at the start.

We secured the premises and asked for the funding of a new lab, named Laboratory of Quantitative Economics (LQE) after its intended users, graduate students of said specialty. The design goals were:

- It should provide a place for interaction among students. Thus, each of them would have his or her own desk and personal computer and there would be some facilities to be shared, like a printer, and an area for socializing.
- It would support general work in Statistics and Econometrics, and target areas like Spatial Statistics and Data Mining which, because of their demanding computing requirements, often exceed what can be done with a laptop or modest desktop machine.
- 3. It would have to be run on a very low overhead, without requiring dedicated staff.
- It would be based on free software, so students willing and able to do so could replicate whatever they found useful in their private machines.

The layout of the Laboratory is completely standard and can be seen in Figure 1. A machine is acting as a server providing, among others, authentication, file storage and printing services. A number of PC's (we usually have between one and two dozen machines running) are networked on the same Ethernet bus, currently 100Mbit Ethernet. Everything is connected to

the Internet.

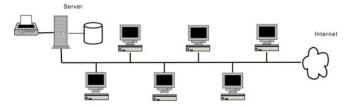


FIG. 1 SKETCH OF ARRANGEMENT AT THE LABORATORY FOR QUANTITATIVE ECONOMICS

The server is a 64bit machine running Linux; we use Debian (http://www.debian.org), and are satisfied with it, but a number of other distributions offer also 64 and 32 bit versions of Linux. What is essential, as we will discuss below, is that *the same* software be available in the server and the client machines.

Students have each his or her own desk and PC. All PC's are fully autonomous machines running 32 bit Debian Linux. However, user files are stored in the server; the PC's mount the relevant directory via NFS, so it appears to be local; this is transparent to the user.

Software installed in both the server and PC's include editors and word processing tools (we encourage the use of Emacs (Stallman, 1997) and LaTeX (Lamport, 1994), but users are free to use the LibreOffice suit or other tools), statistical and econometric software, like R (R Development Core Team, 2008) and Gretel, (http://en.wikipedia.org/wiki/Gretl), among others, GIS tools (Qgis, http://www.qgis.org, GRASS, http://grass.osgeo.org), database tools (PostgreSQL, http://www.postgresql.org; see also Momjian(2001)), PostGIS (http://postgis.net; see also Obe and Hsu (2011)), graphics programs like Gimp (http://www.gimp.org), and an assortment of ofimatics and productivity tools.

The important point is that *the same* software (with very few exceptions, which are not available or are only available in a very different version for 32 and 64 bit machines) is installed in the server and PC's. This brings about two benefits:

- Users can at any point log to the server to start jobs too large for their PC's. The server is a multicore machine each of whose cores runs typically 3 or 4 times faster than a single PC. Since all user files reside in the server, no file movements are involved.
- More importantly, each user is able to log remotely and see exactly the same environment he or she would see in the Lab. All the software and files are there: there is no need to carry anything in pen drives, to upload or download

files.

For the second benefit to be fully realized, it is important that they have suitable means at home. Nowadays, personal machines are ubiquitous, and almost every student has one; most have also high speed Internet access.

While the above setup provides a feasible working environment, we did not reap the full benefits until we added a last and very important piece of software. There are a number of X-terminal software emulators allowing a remote machine to serve as an X-terminal. However, the protocol is fairly verbose, and the amount of information to be transferred fairly voluminous: every keystroke or click requires processing on both the local and remote machine. In our experience, this leads to latency times which are unacceptable for interactive work. Response is much too sluggish when graphical applications are used.

On the other hand, installing these emulators at home (usually on machines running various versions of Windows) was a hurdle for the less computer-savvy users, a problem compounded with the fact that additional layers of software are necessary for the underlying Secure Shell (ssh) protocol.

This problem has been much alleviated by the use of software which compress and streamline the X protocol. Several such pieces of software are in existence using the so-called NX technology. After some experimentation, we have settled for the freely available (although not free source) implementation of NoMachine, available at http://www.nomachine.com.

With this emulator installed, interactive work becomes almost as fast in a remote machine as it is locally. In fact, it is even feasible to stablish a connection and work on the move using a mobile phone or wireless modem. The only problem we have found is that X-terminal emulation is impractical or even impossible with some notebooks supporting only small size / low screen resolution modes.

Administration

It may appear that the setup described shares the problems associated with PC rooms. This is not the case. It is true that each PC has to be installed, but the process can be automated to a large degree, especially if one has control over the network and can have the machines boot from a remote image. Even if you can't, the process of inserting a CD-ROM on each PC and answering a few questions is not overwhelming for small to moderate size labs, as it has to be done only

once. (This can be streamlined; for batches of machines which are equal or very similar, the installation using Linux Debian can be "pre-seeded"; the few questions which one needs to answer at installation can have their answers pre-set.)

Once a minimal installation is working on each machine, everything else, from day-to-day maintenance to installation of new operating system versions or new software, can be done automatically, even for heterogeneous hardware. A tool that we have found invaluable for this purpose is Cfengine (http://cfengine.com; see also Zamboni(2013)). Usually the lab runs for months with no or only minimal need of manual intervention.

Benefits

The LQE has led to a tremendous increase in productivity for students, and almost completely eliminated the wastage of time previously spent in installing software and transporting files, with the following advantages:

- 1. Instant availability, local and remote, of data which is not simple or practical to duplicate: voluminous microdata, digital cartography, etc.
- 2. Increased flexibility in the flow of information. Students, from wherever they happen to be, can spool files directly to their instructors' desk, print them in the lab's dedicated printer or send them in PDF format. This last option has been much used, and affords entirely paper-free interaction. Commenting and/or grading of papers is accomplished with PDF editors (we use Xournal, available at http://xournal. sourceforge.net, but many alternatives, both commercial and open source exist).
- 3. Good interaction among students, while physically at the LQE.

It is true that interaction and cooperative learning are not quite at the same level that the old-style computer lab forced; we estimate that about 80% of the time students work remotely, thus reducing their chances of mutual assistance. We try to enhance student-to-instructor and student-to-student interaction by exploiting forum facilities in e-learning platforms — like Moodle, https://moodle.org, one of the two adopted at our institution.

Our Experience in Perspective

What we have called above "virtual lab" is not without precedents. It is therefore interesting to compare some other experiences to ours and try to ascertain when, how, and why, a virtual lab can add substantially to the learning process.

TABLE 1 TYPES OF LABORATORIES

	Type of access	
Equipment	Local	Remote
Virtual	Local simulation	Virtual Laboratory
Real	Traditional Laboratory	Remote Laboratory

An important matter is to delimit terms. A useful taxonomy is given in Wuttke, Henke and Ludwig (2005), who classify labs according to the sort of equipment used (real, virtual) and access provided (local, remote), as in Table 1.

The equipment is virtual when it mainly consists of software simulators, as opposed to "real" or "hard" equipment (a robot, an oscilloscope, etc.) In the case of "real" equipment, remote access may be severely limited or even impossible: chemicals have to be mixed on the spot, measurement instrument's probes have to be connected, etc.

It is true that complex real systems may be accurately modeled by simulators, and design experiments can be performed and tested remotely. This misses some of the benefits of a real laboratory, replacing reality by a representation of reality. As Keller and Keller(2005) put it,

"...with a simulation you are exploring an algorithm, an abstraction that attempts to model the real world. Any discovery you make is a discovery about the algorithm, not about the world. [...] Rather than using algorithms to create artificial data, one needs to use real experiments..."

Indeed, they go on to say,

"Most technological substitutes for labs are really just computer-based demonstrations that miss the essential elements of science methodology."

It is this lack of reality that is the main drawback of virtual labs in some areas. In the physical sciences in particular the feeling of heat, inertia, strength, speed, light, make a multi-sensorial experience that greatly enhances the learning process. (In Chemistry, smell may also play a role.)

That this is lost in a simulation does not imply that virtual labs are useless. On the contrary, they offer the great benefits: economy, scalability, permanent availability, freeing users from constraints of space and time, even safety: when students have to deal with high voltages, or chemicals involving explosion and inhalation hazards, it is nice to know that the worst

thing it can ever happen is that one needs to press the "reset" button in the computer.

Our subject, Statistics, is somewhat special for the reasons explained. Statistics is Mathematics with an open window on real-world phenomena. Because it is Mathematics, it is ideally suited to computer experimentation which is not tied to a physical location nor does it need measurement instrument's manipulation. And yet it is a subject that benefits from, and indeed *requires*, experimentation. This dual nature of the subject, at once mathematical and "real world", is what it makes it particularly suited for a virtual lab.

What we have called a "dual-faceted virtual lab" spreads over the four cells of Table 1, with the understanding that the the distinction between real and virtual equipment in the rows of Table 1 becomes blurred: the only "real" equipment needed is composed of computers and peripherals, which can be used and controlled remotely just as easily as locally. Aside from computers, the other "input" required is data which, unlike chemicals or oscilloscopes, is immaterial, and can be manipulated with the same ease from a local or remote location. This fact enables us to reap all the benefits of a virtual lab without having to face the "loss of reality" inherent to many other experiences.

Aside from the use of real data, experimentation and teaching in Statistics relies also to a large extent in simulation, but it is doubtful whether the difference data analysis/simulation warrants the Virtual/Real distinction in Table 1.

At the time of writing, the LQE has been in operation for over five years, and is therefore a well established experience, well beyond a proof of concept.

Experiences similar to ours target energy efficiency, (NComputing, Inc., 2010), enhancing active and cooperative learning, (Moor, 2006), or web-based arrangements (see (Jalobeanu, 2006) and references therein). Closer to our design goals is (Nixon and Dwolatzky, 2002), although they mainly address (back in 2002) presential labs.

In the field of Statistics, most experiences we have been able to track address mainly the simulation side, providing elaborated Java applets to demonstrate (and help visualize) different statistical concepts, that are otherwise difficult to grasp for students. They would fall in the "Virtual Laboratories" cell of Table 1.

In some instances they also include case studies, sophisticated multimedia teaching material and online texts. Notable examples are the Rice Virtual Lab in Statistics (http://onlinestatbook.com/rvls.html), or the University of Alabama's Virtual Laboratories in Probability and Statistics (http://www.math.uah.edu/stat/). Both these sites offer great educational material online, mostly simulation-oriented. Their emphasis differs from ours in that they do not seek to integrate a real and remote lab.

Not specifically oriented to Statistics, but also covering that field, is the Virtual Lab of the University of North Carolina at Chapel Hill, (http://its.unc.edu). Unlike our goal of seamless integration of local and remote computing, their emphasis appears to be in granting access to University-licensed applications on a wide variety of students-owned platforms, including Windows, Linux, Unix and Android clients. Their Virtual Lab is based on Cytrix Technology virtualization software and relies, so much as Statistics is concerned, in commercial applications.

Printing and storage resources can also be shared, so that each remote client can access network as well as local storage; this provides the same or even greater functionality than we have implemented. Unlike in our approach, students do not see by default a common interface and environment from wherever they happen to work; but everything can be shared and work done locally at the university can be continued remotely.

Our experience shows that a dual-faceted lab, offering both local and remote resources, is a feasible alternative, and relatively cheap to build with off-theshelf hardware and (mostly) free software. Its success in our case has been critically dependent on:

- Being able to offer a user interface that looks almost exactly the same, whether locally or remotely;
- The general availability to students of hardware and wide band connections, and
- A simple, "click-and-install" piece of software that handles remote connections efficiently.

The downside, as mentioned before, is reduced student interaction when they work remotely, but use of fora (and, in the future, chats), may go a long way to alleviate that. Lack of student interaction is particularly worrisome in that it prevents mutual help in simple, procedural tasks ("How do you merge these two files in different formats?" "What does this error message mean?").

The Road Ahead

We have toyed with the idea of extending the integration right to the student's laptops. One can envision the idea of wireless synchronization of the student's laptops as they switch them on in the LQE premises. The centralized automatic administration setup with Cfengine as the cornerstone, would make that possible. In fact, we have taken advantage of the same tool to manage also some machines in our Department, alien to the LQE.

We are still searching for ways which would enable our students to conduct work in their own machines, while not in the Lab and away from an Internet connection, and yet enjoy the maximum of the facilities offered by the Lab, with automatic, or user controlled, synchronization of data files as soon as they eventually regain Internet connectivity.

We would like students to be able to pursue their work in their machines, with a familiar environment, irrespective of whether they are connected to the Internet, and be able to integrate easily their work with their LQE-based account. Sophisticated users can achieve something along these lines, at least for flat text files, using version control systems such as git (http://git-scm.com; Chacon, S. (2009)), but this is much too complex (and requires far too much manual interaction) for the vast majority of students.

Even though this vision of seamlessly integrating student's machines would be a giant step forward, difficulties are daunting: student's computers are widely different in capabilities, and enforcing the few necessary rules might be an administrative nightmare. Oliver, Mollette and Corn (2012) report one case in which providing file storage synchronization for students and teachers, which is only part of what we envision, had to be discontinued.

Conclusions

An experience has been presented whose goal is to provide a seamless environment for statistical work both in local facilities and remote connections. We have shown how using (mostly) open source software standard hardware and well-established, mature and open standard technologies, this can be achieved with a very low budget and without committing large resources of human power.

We also have described the particular combination of factors that have made our experience successful: unlike in most natural sciences, statistical work can be done remotely just as well as locally, there is general availability of computer hardware and fast Internet connections among our students, and very high quality open source software exists for all our needs.

Most importantly, the learning aspects have been addressed. Among them, the most important drawback detected is the reduced student interaction. In this connection, the availability of local premises for the students to use appears to us of the utmost importance, even if (as in our case) little use is made of them as compared to remote connections. Although relatively little used, the local facilities offered by the LQE afford a measure of student and student-teacher interaction which e-learning platforms, chats and e-mail can only partly replace.

ACKNOWLEDGMENTS

Funding from grants EOPT (IT-567-13, Basque Government), BETS (University of the Basque Country, UPV/EHU) and MTM2012-31514 (Ministry of Economy and Competitiveness, Spain) is gratefully acknowledged.

REFERENCES

Chacon, S. (2009) Pro Git, APress, 2009.

Jalobeanu, M. (2006). Internet - a virtual laboratory for distributed computing. In Auer, M. E. and Ursutiu, D., editors, 2006 International Remote Engineering Virtual Instrumentation Symposium.

Keller, H.E. and Keller, E.E. (2005) Making Real Virtual Labs, The Science Education Review, 4(1), 2-11.

Krafft, M. F. (2005). The Debian System. No Starch Press.

Lamport, L. (1994). A Document Preparation System. LATEX User's Guide and Reference Manual. Addison-Wesley.

NComputing, Inc. (2010). Santa Ana College creates green lab with virtual desktops. Retrieved from http://www.ncomputing.com/filedepot_download/6478/264

Momjian, B. (2001). *PostgreSQL: Introduction and Concepts*. Addison-Wesley

Moor, S.S. (2006). Case study: Renovating a computer teaching laboratory for active and cooperative learning. In *Proceedings of the 2006 Illinois-Indiana and North Central Joint Section Conference*. American Society for Engineering Education.

Neteler, M. and Mitasova, H. (2007). *Open Source GIS: A GRASS GIS Approach*. Springer Verlag.

Nixon, K. and Dwolatzky, B. (2002). Computer laboratory

infrastructure in engineering education- a case study at Wits. In Africon Conference in Africa, 2002. IEEE AFRICON. 6th, volume 1, pages 431 – 436.

Obe, R.O. and Hsu, L.S. (2011) *PostGIS in Action*, Manning, 2011.

Oliver, K.M., Mollette, M. and Corn, J. (2012) Administrative Perspectives on the Implementation of One-to-One Computing, J. of Information Technology and Application in Education, vol. 1(4), 125—142.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ritchie, F. (2008) Secure access to confidential microdata: four years of the Virtual Microdata Laboratory, Economic & Labour Market Review, vol. 2(5), 29-34.

Stallman, R. M. (1997). *GNU Emacs Manual*. Free Software Foundation.

Wuttke, H-D., Henke, K. and Ludwig, N. (2005) Remote Labs versus Virtual Labs for Teaching Digital Systems Design, International Conference on Computer Systems and Technologies – CompSysTech' 2005.

Zamboni, D. (2013) *Learning CFEngine 3*, O'Reilly, Sebastopol, CA.



Araceli Garín, PhD. Mathematics (UPV/EHU, 1992) She is an Associate Professor at the School of Economics and Management, UPV/EHU at Bilbao (Spain). Her main interests lie at Mathematical Programming with and without uncertainty in the parameters.

She has published papers in Computers and Operations Research, European Journal of Operational Research or Annals of Operations Research, among some others.

Dr. Garín leads the EOPT, a group doing research in Statistics and Optimization. In cooperation with similar groups in Spain and Norway, they are involved in algorithmic development and applications of stochastic optimization in Finance and the oil industry.



Fernando Tusell, PhD. Economics (UPV/EHU, 1983). M.Sc. (Math) (MIT, 1983).

He is a Professor at the School of Economics and Management, UPV/EHU at Bilbao (Spain). His main interests lie at the intersection of Statistics and Computing. He has published papers in Computational Statistics and Data Analysis, Statistics and Computing and Journal of Statistical Software.

Prof. Tusell has previously served as statistician in local government and in the European Comission.



Aitziber Unzueta, PhD. Banking and Quantitative Finance (UPV/EHU, 2012), M.Sc. Mathematics (UPV/EHU, 2006)

She is a Lecturer at the Technical School of

Engineering, UPV/EHU at Bilbao (Spain). Her interests lie at Algorithmic Development for large scale optimization problems. She has published papers in Computers and Operations Research and Top.

Dr. Unzueta has been associated with the EOPT, specializing in Stochastic Optimization. While writing her thesis she visited the Research Center CARISMA, Brunel University, United Kingdom.