INTERNATIONAL JOURNAL OF **DIGITAL AND DATA LAW**

REVUE INTERNATIONALE DE DROIT **DES DONNÉES ET DU NUMÉRIQUE**

Vol. 3 - 2017

ISSN 2553-6893

IMODEV
LES ÉDITIONS

International Journal of Digital and Data Law
Revue internationale de droit des données et du numérique

Direction :
Irène Bouhadana & William Gilles

ISSN : 2553-6893

IMODEV
49 rue Brancion 75015 Paris – France
www.imodev.org
ojs.imodev.org

Les propos publiés dans cet article
n'engagent que leur auteur.

The statements published in this article
are the sole responsibility of the author.

À PROPOS DE NOUS

La **Revue Internationale de droit des données et du numérique (RIDDN)/ the International Journal of Digital and Data Law** est une revue universitaire créée et dirigée par Irène Bouhadana et William Gilles au sein de l'IMODEV, l'Institut du Monde et du Développement pour la Bonne Gouvernance publique.

Irène Bouhadana, docteur en droit, est maître de conférences en droit du numérique et droit des gouvernements ouverts à l'Université Paris 1 Panthéon-Sorbonne où elle dirige le master Droit des données, des administrations numériques et des gouvernements ouverts au sein de l'École de droit de la Sorbonne. Elle est membre de l'Institut de recherche juridique de la Sorbonne (IRJS). Elle est aussi fondatrice et Secrétaire générale de l'IMODEV.

William Gilles, docteur en droit, est maître de conférences (HDR) en droit du numérique et en droit des gouvernements ouverts, habilité à diriger les recherches, à l'Université Paris 1 Panthéon-Sorbonne où il dirige le master Droit des données, des administrations numériques et des gouvernements ouverts. Il est membre de l'Institut de recherche juridique de la Sorbonne (IRJS). Il est aussi fondateur et Président de l'IMODEV.

IMODEV est une organisation scientifique internationale, indépendante et à but non lucratif créée en 2009 qui agit pour la promotion de la bonne gouvernance publique dans le cadre de la société de l'information et du numérique. Ce réseau rassemble des experts et des chercheurs du monde entier qui par leurs travaux et leurs actions contribuent à une meilleure connaissance et appréhension de la société numérique au niveau local, national ou international en en analysant d'une part, les actions des pouvoirs publics dans le cadre de la régulation de la société des données et de l'économie numérique et d'autre part, les modalités de mise en œuvre des politiques publiques numériques au sein des administrations publiques et des gouvernements ouverts.

IMODEV organise régulièrement des colloques sur ces thématiques, et notamment chaque année en novembre les *Journées universitaires sur les enjeux des gouvernements ouverts et du numérique / Academic days on open government and digital issues*, dont les sessions sont publiées en ligne [ISSN : 2553-6931].

IMODEV publie deux revues disponibles en open source (ojs.imodev.org) afin de promouvoir une science ouverte sous licence Creative commons_CC-**BY-NC-ND** :

1) la *Revue Internationale des Gouvernements ouverts (RIGO)/ International Journal of Open Governments* [ISSN 2553-6869] ;

2) la *Revue internationale de droit des données et du numérique (RIDDN)/International Journal of Digital and Data Law* [ISSN 2553-6893].

About us

The **International Journal of Digital and Data Law / Revue Internationale de droit des données et du numérique (RIDDN)** is an academic journal created and edited by Irène Bouhadana and William Gilles at IMODEV, the Institut du monde et du développement pour la bonne gouvernance publique.

Irène Bouhadana, PhD in Law, is an Associate professor in digital law and open government law at the University of Paris 1 Panthéon-Sorbonne, where she is the director of the master's degree in data law, digital administrations, and open governments at the Sorbonne Law School. She is a member of the Institut de recherche juridique de la Sorbonne (IRJS). She is also the founder and Secretary General of IMODEV.

William Gilles, PhD in Law, is an Associate professor (HDR) in digital law and open government law at the University of Paris 1 Panthéon-Sorbonne, where he is the director of the master's degree in data law, digital administration and open government. He is a member of the Institut de recherche juridique de la Sorbonne (IRJS). He is also founder and President of IMODEV.

IMODEV is an international, independent, non-profit scientific organization created in 2009 that promotes good public governance in the context of the information and digital society. This network brings together experts and researchers from around the world who, through their work and actions, contribute to a better knowledge and understanding of the digital society at the local, national or international level by analyzing, on the one hand, the actions of public authorities in the context of the regulation of the data society and the digital economy and, on the other hand, the ways in which digital public policies are implemented within public administrations and open governments.

IMODEV regularly organizes conferences and symposiums on these topics, and in particular every year in November the Academic days on open government and digital issues, whose sessions are published online [ISSN: 2553-6931].

IMODEV publishes two academic journals available in open source at ojs.imodev.org to promote open science under the Creative commons license CC-**BY-NC-ND**:

1) the *International Journal of Open Governments*/ la *Revue Internationale des Gouvernements ouverts (RIGO)* [ISSN 2553-6869] ;

2) the *International Journal of Digital and Data Law* / la *Revue internationale de droit des données et du numérique* (RIDDN) [ISSN 2553-6893].

SOME THOUGHTS ON THE USE OF STATISTICAL SAMPLING IN LEGAL RESEARCH

by **Carlos N. BOUZA-HERRERA**, Professor at the Faculty of Mathematic of University of Havana, Cuba.

Much of legal research is based on discovering facts through analyzing a lot of papers. Electronically Stored Information (ESI) poses issues on using data stored electronically. With the increase of data volumes, a need of reducing costs, without violating the accepted assumptions poses urgently mid changes in the law firms. The reduction of costs should not be solved by using "low-cost lawyers".

This paper discusses on the use of Technology Assisted Review and Statistical Sampling for retrieving information and some examples are discussed for illustrating.

A broad definition of legal research is that: it is a process which looks for identifying and retrieving what is needed for supporting legal decision-making. Hence, we may consider that it starts with the analysis of the facts on a particular and ends with the binomial application-communication of the results of the investigation.

Nowadays statistical evidence, sustained by probabilistic reasoning, plays an important role in common life. It is expanding its area of applications to criminal investigations, prosecutions and trials. Particularly, forensic scientific evidence, including DNA, produced by expert witnesses, is one of the emerging areas for statistical applications. That sustains that if you are involved in criminal adjudication, having a comprehension of the basics of probability and statistics is needed. In other legal researches, a similar situation is present: data must be retrieved and analyzed. Misunderstandings of what statistical information at hand are to be processed and interpreted, as well as of the role of the involved probabilities, have contributed towards serious miscarriages of justice. These facts suggest including in the education for lawyers a training on statistical thinking on how it should be used in legal research.

Actually, some processes use statistical sampling for providing evidence at the court yard. The correctness of the statistical procedures used, are being taken into account in the allegation of decisions by the court. Hence, having a good statistic advisor is one of the actual needs of the law firms.

Another problem is related with the need of dealing with Bigdata. They are being used in different legal issues at least in the past 20 years. The presence of Bigdata poses to the investigators to deal with responding to:
– How much data they have?
– Which is the structure of the data (structured, unstructured, text-based, internal and external)?
– Is it possible to analyze the existing data in real time for instantaneous decision-making?
– Are the data reliable?
Much of legal research is based on discovering facts through analyzing a lot of papers. Electronically Stored Information (ESI) poses issues on using data stored electronically. With the increase of data volumes, a need of reducing costs, without violating the accepted assumptions poses urgently mid changes in the law firms. The reduction of costs should not be solved by using "low-cost lawyers".

To give a modern response when dealing with Bigdata an emerging technology for the retrieval of document information is connected with Technology Assisted Review (TAR) and Statistical Sampling. They are occupying a distinguished place as a tool for the research of law firms, as it reduces risks and improves productivity in eDiscovery processes. TAR is based on statistical models and it is being accepted as some kind of standard statistical tool for analyzing Bigdata problems posed by the existence of ESI.

Actually, many U.S. courts are endorsing the use of predictive-coding technologies. Consequently U.S. law-firms are encouraging structuring task groups for improving Bigdata practice.

§ 1 – Some uses of statistical sampling

The analysis of data always has posed a complicated task to law firms. Nowadays the available data overcomes the capacity of the attorneys if some modern technique is not used for sampling and providing relevant information. Consider the use of applying Statistical Sampling to discovery of relevant and responsive documents. Though it is not a common practice, it is increasing its role in legal research. The reasons are the usual in statistical research. Its use is cost-effective in many te sts as it s behavior has been reasonably effective in finding relevant and responsive documents.

Sampling is currently used in many areas of the Social Sciences. In particular, sociology studies use sampling models for obtaining information. The theoretical frame uses the fact that the study deals with finite population of well identified units ($U=\{u_1 ,\ldots,u_N\}$).

Selecting a sub set of them generates a sample (s⊂U). Using judgmental sample was initially the approach up to the general acceptation that probability sampling is the only way of obtaining "representative samples".

Statistical sampling is of use in many aspects of the administration of justice. Providing facts coming from well supported statistical research is a source of evidence. The court analyzes the results of statistical research but it must be aware of what is scientifically correct and how some models may be used for manipulating the results. Then, statistical experts are to be contracted by law firms for designing their needs of developing statistical inquires. On the other hand, the court must have an adequate counterpart for giving support to the righteous of the conclusions derived by the research.

The use of statistical evidence has proved to be of considerable support in the court. In some areas, they are currently used.

Statistical sampling is accepted for estimating Medicare overpayments. Unfortunately, there are not well-established guidelines for sampling methodologies, as in other areas. Hence, there is a basis for considering whether a statistical principle, or method, is to be preferred to another one. There is a need of establishing some standards for considering when a statistical study is valid or not. In USA, the programs of Medicare have established that a statistical sampling evidence should be considered as acceptable, only if it uses a probability sampling design. That is observing any sample s must have a probability P(s) ∈ [0,1].

The importance of modeling adequately is exemplified by some trials as the following ones:

– *Transyd Enterprises LLC D/B/A Transpro Medical Transport (Appellant) vs (Beneficiaries) Trailblazer Health Enterprises LLC (Contractor)*, Claim for Part B. Benefits, 2009 WL 5764287 (Sept. 15, 2009). MAC rejected the appellant's argument "PSC's sampling methodology is invalid" because the PSC failed to document that its statisticians possessed at least a master's degree in statistics or the equivalent.

– *Robert D. Lesser, M.D. & Assocs. (Appellant) vs (Beneficiaries) Pinnacle Business Solutions, Inc. (Contractor)*, Claim for Part B Benefits, 2011 WL 5263619, Docket No. M-11-358 (Feb. 18, 2011). The Council noted that ALJ relied on the 60-day timeline in the MPIM, which applies to prepayment and post payment review for MR (Medical Review) purposes. The case arose from a statistical sampling review by the Benefit Integrity unit of the ZPIC.

– *The MMPIM General Medicine, P.C. (Appellant) (Beneficiaries) Palmetto, GBA (Contractor)*, Claim for Part A Benefits, 2010 WL 7232825, Docket No. M-10-1933 (Nov. 24, 2010): The Council found appellant's case was based on unsupported speculations and conjectures. It addresses claimed that stratification should have been used, stating that the statistical sampling guidelines did not require stratification of every sample in order to make the sampling valid.

§ 2 – Technology Assisted Review (TAR) and Statistical Sampling

A particularly important task in legal work is text classification. Different studies suggest that machine learning techniques outperforms the classic manual document review developed by lawyers. They support that Technology Assisted Review (TAR) and Statistical Sampling increase both productivity and accuracy at a lower cost. Empirical evidence sustains that the use of TAR reduces the review time in a 75% of the time and the cost is only 30% of the classic methods.

Those are the reasons why one of the more accepted sampling procedures is using TAR. There are not many publications on its theoretical properties but the comparison of the cost reduction, due to its use has increased its popularity in legal research. Many law firms are considering how unassisted document review performs in comparison with TAR, which is validated by statistical sampling models.

TAR uses the expertise of attorneys and the methods of machine-learning to automatize the prioritization of documents to be reviewed. The ranking uses a measure of the responsiveness of document to a particular matter. By using it for dealing with Big data the firms reduce costs and key documents are obtained faster.

Some recent documented evidences of the usefulness of using TAR are:

– *Da Silva Moore vs Publicis Groupe8*. Andrew Peck (US Magistrate Judge) gave his opinions the validity of judgmental and statistical sampling for validating the results of predictive coding. (The Case for Statistical Sampling in e□Discovery7).

– *Kleen Products vs Packaging Corporation of America9*. Nan Nolan (US Magistrate Judge) heard the testimony, for sustaining the validity of the sampling process, used by defense. The validity of testing the results based on research terms, instead of predictive coding, in finding relevant documents was on trial. the parties had to determine with sampling procedure was acceptable for them. Once an agreement was obtained on the keyword to be looked for using sampling the research and discussion went forward.

– *Global Aerospace vs Landow Aviation*. The court stated that predictive coding (aka TAR) including a statistical model for validating the protocol was adequate for locating and retrieving documents for production.

The reduction of the costs is important but in addition the consistency of using TAR and sampling is considerably larger than the so called "linear review". Linear reviews are developed by the reviews, performed by attorneys of the documents. The inconsistency of the reviews due to human error in not measured. Commonly no statistical sampling is used and hence, the reliability of such reviews is not possible. Therefore, the inconsistency of reviewers is unknown. TAR is validated with statistical sampling and it is highly consistent, and hence more reliable, compared with results of unassisted reviews performed by attorneys. Therefore, using it the lawyers assure that the process achieves a large level of success in identifying the relevant and responsive documents.

Well known sampling models as stratification allows improving the quality of the review process. For example, if a ranking of the importance of the documents is made previously, the consistency may be improved by using an unequal probability sampling or ranked set sampling. Such approaches save time as they avoid expecting for "first-level" reviews. For example, documents ranked first receive a preferential treatment in terms of the probability of being selected.

Corporate law departments deal with large amounts of data from invoices, and need to determine the factors influencing rates for negotiating better, deals based on that data. a free mobile application that aggregates data from thousands of law firm invoices is TyMetrix Legal Analytics. TyMetrix RateDriver™ mobile application uses the statistical model from Real Rate Report™. It is a statistical analysis of legal invoices.

§ 3 – A STUDY

Less documented is its use in providing evidence on reclamations on the contamination due to enterprises. A question is: are the levels of contamination acceptable? The enterprise produces reports to the governmental agencies. On doubts on the accuracy of the reports environmentalists supported claims of farmers that the water used for agriculture was being contaminated. Their claim is based on the observed behavior of the production of the land.

The case was *Farmers, F. (Appellant) vs (Beneficiaries) Chemical enterprise, CE (Contractor)*, Claim for Part A contamination of the water is affecting the fertility of lands: The appellant considered that the reported data which supported that the contamination levels were within the accepted interval were not correct. The arguments of the enterprise were unsupported speculations and the conjectures cannot be proved without a statistical study. The statisticians supporting the appellants claimed that the measurements of the sensors at the factory output were providing not accurate information. They selected some points in the course of the water source and obtained their own measurements. A sample of them were compared with the ones made at the output of the factory by the sensor of the enterprise owners.

The set of measures of the outputs were considered as binary (0, 1) indicating whether they coincided with the ones of the other sensors (correctly classified=1, incorrectly classified=0). The results of N measurements are summarized in the Table 1.

Considering that the classification is equivalent to a double-blind method that is they are made independently. Each measurement generates a value

$$Y_t = \begin{cases} i \; if \; t \; is \; correctly \; classified \; by \; both \; sensors \\ 0 \; other \; wise \end{cases}$$

Summarizing is obtained the next table

Table 1. Classification of N measurements of 2 sensors

	Correct	**Incorrect**	**Total**
Correct	$n_{11} = A$	$n_{12} = B$	$n_{1+} = A + B$
Incorrect	$n_{21} = C$	$n_{22} = D$	$n_{2+} = C + D$
Total	$n_{+1} = A + C$	$n_{+2} = B + D$	$N = A + B + C + D$

Different agreement indexes were considered. They are function of

$$p_{ij} = \frac{n_{ij}}{n}, \qquad q_2 = \frac{1}{2}(p_{2+} + p_{+2}), \qquad q_1 = \frac{1}{2}(p_{+1} + p_{1+}),$$

where $$p_{i+} = \frac{n_{i+}}{2}, \; p_{+i} = \frac{n_{+i}}{n}$$

Were evaluated the following indexes

Dice

$$I_D = \frac{2p_{11}}{p_{1+} + p_{+1}} = \frac{p_{11}}{p}$$

A value close to zero means that the sensors have a small "agreement".

Correlation coefficient

Note that we are dealing with attributes (categorical variables). In this case, the correlation coefficient of Pearson may be rewritten, in terms of Table 1 as

$$\rho \quad = \frac{A \times D - B \times C}{\sqrt{(A+B) \times (C+D) \times (A+C) \times (B+D)}}$$

The values of ρ will be in the interval [-1, 1]. If $\rho \approx 1$, the sensors behave similarly, $\rho < 0$ means that they highly disagree and they are “independent” if $\rho \approx 0$.

Measure of Differences

$$D = \frac{B + C}{N}$$

An increase of D means that they largely disagree

Kappa

$$\kappa = \frac{\sum_{i=1}^{k} p_{ii} - \sum_{i=1}^{k} p_{i+}p_{+i}}{1 - \sum_{i=1}^{k} p_{i+}p_{+i}}$$

A large value of it means the existence of a high level of agreement.

3 sensors were placed and data were collected during a month. The values of the indexes were computed for each one and compared with the reports of the enterprise. Each one was evaluated considering the belonging to the accepted levels of contamination fixed by the law. They are reported in the next table

Table 2. Values of the indexes of 3 sensors

Sensor	Dice	Correlation coefficient	Measure of Differences	Kappa
1	0,801	0,128	0,333	0,302
2	0,823	-0.001	0,301	-0,010
3	0,774	0,300	0,352	0,364

Then, it was documented that the lectures of the enterprise had a low agreement when classifying the violating of the accepted level of contaminator with the other sensors.

The court fixed a fine to the enterprise for avoiding their responsibilities with the environment and a calculation of the damage to the farmers is in progress. The statisticians of the enterprise alleged that they assumed that the measurements were normally distributed but the appellant´s proved that this probability assumption was incorrect and that categorical data analysis must have been used by then for controlling.

References

AITKEN C., P. ROBERTS and G. JACKSON, *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings. Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*, Royal Statistical Society's Working Group on Statistics and the Law, 2010.

AITKEN C.G.G. and F. TARONI, *Statistics and the Evaluation of Evidence for Forensic Scientists. Chichester*, Wiley, 2004.

ALLEN R.J. and M. PARDO, "The Problematic Value of Mathematical Models of Evidence", 36 *Journal of Legal Studies* 107.

BALDING D.J., *Weight-of-Evidence for Forensic DNA Profiles*, Chichester, Wiley, 2005

BARON J.R., "Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search", *XVII Rich. J.L. & Tech. 9,* 2011: http://jolt.richmond.edu/v17i3/article9.pdf.

DEGROOT M. H., S.E. FIENBERG, and J.B. KADANE, (eds.), *Statistics and the Law*, New York, Wiley, 1994.

HODGSON D., "Probability: The Logic of the Law – A Response", 15 *Oxford Journal of Legal Studies* 51, 1995.

KADANE J.B., *Statistics in the Law: A Practitioner's Guide, Cases, and Materials*, New York, OUP, 2008

KOEHLER J.J., M.J. SAKS and J.J. KOEHLER, "The Coming Paradigm Shift in Forensic Identification Science", 309 *Science* 892, 2005

PASKACH C. H., F. E. NELSONAND, M. SCHWAB, "The Case for Technology Assisted Review and Statistical Sampling in Discovery", *DESI VI Workshop*, ICAIL Conference, San Diego, CA, 2015

THE CLARO GROUP. L.L.C., W.C. THOMPSON and E.L. SCHUMANN, "Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy", 11 *Law and Human Behaviour* 167, 1987

SHARP M., "Text Mining", Rutgers University, School of Communication, Information and Library Studies, 2009: http://www.scils.rutgers.edu /~msharp/text_mining.htm. [Accessed: september, 2016]