Institution: City, University of London

Unit of Assessment: 11 Computer Science and Informatics

Title of case study: Speeding up familial DNA searching using Bayesian networks

Period when the underpinning research was undertaken: 2003 - 2015

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Dr Robert Cowell	Associate Professor of Statistics	02/10/1995 - 30/06/2016

Period when the claimed impact occurred: 2005 - 2020

Is this case study continued from a case study submitted in 2014? Yes

1. Summary of the impact (indicative maximum 100 words)

This case study provides an account of work on the application of Bayesian networks to problems of interest in forensic genetics. Developed in research at City, University of London an Expert System-based software programme has dramatically improved the speed, efficiency and cost-effectiveness of forensic laboratory methods used in familial DNA analysis and in criminal investigation. The software has been in continuous use by the UK's leading forensic supplier, Eurofins plc, to analyse DNA evidence for forensic identification in determining suspects in violent crimes, solving paternity/maternity, in cases of missing persons and identifying human remains of victims from mass disasters including at Grenfell Tower. The software is among new technical methods that can be used on the surviving evidence of 'cold cases', often with conclusive results.

2. Underpinning research (indicative maximum 500 words)

Dr Robert Cowell has spent over 20 years carrying out joint research with Professor Steffen Lauritzen FRS currently Professor of Statistics at the University of Copenhagen and Professor Julia Mortera of Universita Roma Tre, Italy on the modelling and analysis of DNA mixture samples that are recovered from crime scenes. Given one or more such DNA samples, two questions of particular interest that their work addresses are: (1) given the DNA profile of a possible suspect, what is the weight of evidence that the suspect contributed their DNA to the sample?; and (2) if no suspect is at hand, can we obtain the genetic profile or profiles of the unknown contributors to the crime sample, for use for example in a search of a DNA database of previously convicted individuals to look for a DNA match? An extension of the second question is familial DNA searching which involves the search of a DNA database to detect and statistically rank a list of potential candidates in the DNA database who may be close biological relatives (e.g., parent, child, sibling) to the unknown individual contributing the evidence DNA profile, combined with lineage testing to help confirm or refute biological relatedness. [3.1],[3.2]

The determination of the likelihood that two or more individuals are biologically related, based on their DNA profiles is of interest in both civil and criminal application. A common civil application occurs in a case of disputed paternity in which, for example, a mother claims that the father of her child is a certain male, but he denies this. A related criminal application is that of incest, for example when a man is suspected of fathering a child by his daughter. The establishment of relatedness of individuals can also be important for immigration cases. The use of Probabilistic Expert Systems (PESs), also called Bayesian networks in familial DNA searching has great potential in these cases, however, general purpose PES software is not particularly well-suited to the repetitive tasks of specifying an appropriate set of marker networks for a specific problem, editing the many local conditional probability tables and combining evidence from several

genetic markers to evaluate likelihoods. Such software can be time-consuming and error-prone because of the number and sizes of the tables requiring specification in the Bayesian networks. [3.2]

The creation and development of a novel prototype computer programme called FINEX overcame these problems [3.3]. FINEX was originally written to automate the process of constructing Bayesian networks PESs and to provide a user-friendly interface by reproducing the usual genetic tree in the computer screen. Thus, the program shields the user from the technicalities and tedium of specifying Bayesian networks directly, eliminating errors that could arise in specifying the conditional probability tables. The Bayesian network is used to structure a definite genetic problem (in our case, a disputed relationship) in terms of a graphical model (with elementary deterministic relations, probabilistic computational nodes and a query node). Later, FINEX was extended to carry out the calculations itself, without having to export the Bayesian networks to a separate piece of software for analysis.

FINEX allows a user to express the structure of a forensic identification problem in a quick and simple manner through the syntax of a high-level graphical specification language. This allows quite complex hypotheses to be entertained regarding the relationships of individuals which could be so complex that an expert forensic scientist could not do the calculations. The user of the programme specifies two or more hypothetical relationships and the software evaluates the likelihood of the hypothetical relationships between known genetic profiles being actual. Assessments are made based on the differences of the likelihoods of the hypotheses.

It is the speed of the program for carrying out the probability calculations, approximately 2000 individual profiles per minute, that makes a large-scale search of a database possible in a reasonable time. The algorithms by which FINEX converts the user input in the graphical specification language and data on observed markers to the Bayesian networks used in PES are described in research output.

Another genetic-related area of research is that of learning the pedigree, or family tree, of a group of closely related individuals given the DNA proles of the individuals. This can have a number of applications. It is of interest for biologists studying animal populations. In the human domain there are two of particular interest. One is of identifying family groups in a mass grave, the other is that of identifying deceased victims in a mass disaster even such as a plane crash by comparing their DNA with that of living relatives. The result of the *Theoretical Population Biology* paper [3.4] presents an algorithm for carrying out an exhaustive search for up to 30 individuals, and software that I developed for this has been incorporated into the public domain FRANz pedigree reconstruction software available from the University of Leipzig [3.5].

Cowell works at the interface of Artificial Intelligence and Statistics and has worked on topics such as decision theory, statistical inference, Bayesian networks, graphical models and expert systems. His main area of research for the last twenty years has focussed on the interface of probabilistic graphical models, machine learning, and artificial intelligence, in particular in the theory and application of Bayesian Networks. These have come to high prominence because of their flexibility, widespread applicability and computational efficiency.

3. References to the research (indicative maximum of six references)

3.1 Cowell, R.G., Lauritzen, S.L. and Mortera, J. (2007a). A gamma model for DNA mixture analyses, *Bayesian Analysis*, **2**(2), 333–48

3.2 Cowell, R.G., Lauritzen S. L. and Mortera J. (2011), 'Probabilistic expert systems for handling artifacts in complex DNA mixtures', *Forensic Science International: Genetics*, **5**(3), p.202-209

3.3 Cowell R.G. (2003). FINEX: a Probabilistic Expert System for forensic identification, *Forensic Science International,* **134**(2-3), p.196-206

3.4 Cowell, R.G. (2009) A simple greedy algorithm for reconstructing pedigrees, *Theoretical Population Biology*, **83** p.55-63 (February)

3.5 FRANz Beta Pedigree Reconstruction: A fast and flexible parentage inference program for natural populations. available for free download [Information retrieved 11 May 2020]

Indicators of quality for the underpinning research:

Research was supported by a The Leverhulme Trust Interchange Research Grant: *Bayesian networks for forensic inference from genetic marker*, October 2001 to September 2004. Principal Investigator: A. Philip Dawid FRS then at University College London. Funding value: £98,330 Grant number: F/071134/K

Dr Cowell was awarded the 2001 DeGroot Prize by the International Society for Bayesian Analysis along with A. Philip Dawid FRS, Steffen Lauritzen FRS, and Sir David J. Spiegelhalter FRS 'for research concerned with fundamental issues of statistical inference, decision theory and/or statistical applications.'

4. Details of the impact (indicative maximum 750 words)

FINEX software was introduced to the DNA Unit of the UK Forensic Science Service (UK FSS) in 2006 and was used regularly in criminal casework. Following the closure of the Forensic Science Service in 2012/3 the software was licenced to VidaVia Media SL which has collaborated in a further re-branding of the application as **GPS-ibd**, a trademark of Gene Pool Systems (a trading name of VidaVia Media SL). "ibd" signifies "Identical By Descent". The software has been licenced to three users in this REF period: United States Department of Defence, LGC Forensics, and ForGenetica Consultants Ltd which has carried out further functional and casework testing and support for commercial purchases of GPS-ibd. [5.1], [5.2]

GPS-ibd provides fast throughput of cases and decisions and is used in standard relationship forensic casework for use in routine civil and court ordered paternity tests, relationship analysis for criminal casework (including rape and incest cases), coronial work, missing persons and mass fatality cases, and commercial contract relationship analysis for European and Middle Eastern clients. In addition, an advanced processing functionality of GPS-ibd allows the application to be used for Familial Searching of the Police/National DNA database. [5.3]

LGC plc, the largest player in UK's forensics market since the closure of UK-FSS reported that the product was "*successful*" and "*Ground breaking*". LGC plc has used the software continuously on "*some of the most high-profile cases in the past 2 years (including cases that have been highly publicised, highly political and deeply tragic)".* [5.4]

In March 2016, LGC was successful in gaining IS0/IEC17025:2005 accreditation through the United Kingdom Accreditation Service, under schedule no. 0003 for Relationship Analysis Services using GPS-ibd software, effective from 12 February 2016. ISO/IEC 17025:2005 specifies the general requirements for the competence to carry out tests and/or calibrations, including sampling. It covers testing and calibration performed using standard methods, non-standard methods, and laboratory-developed methods [5.5]. ISO accreditation is an important milestone in product development. It gives assurance to potential clients and makes the results less open to challenge by the courts.

Commenting on the introduction of the new GPS-ibd service following its accreditation, Dr Tim Clayton, said "*The GPS-ibd software has significantly increased our capabilities to deal with the more complex pedigrees that can be encountered in forensic casework and allows us to provide a more comprehensive service to our customers. In addition, as the software does not require expensive servers or any other system architecture, it can be deployed on a desktop and runs calculations within seconds. The automation of complex calculations means that computation is less error prone and far more efficient*." [5.5]

Cases where GPS-ibd was used:

In 2017, pharma and life sciences company Eurofins Scientific acquired the forensics division of LGC, the largest player in UK's forensics market. Along with the transfer of the trade of that division, 706 employees and the licence for GPS-ibd were transferred to Eurofins Scientific. According to a Senior Reporting Scientist – DNA Science Lead at Eurofins Forensic Services the software "*is used almost every day*" and that it "*has been used in hundreds of criminal cases involving kinship issues, most commonly rape*". [5.6]

Immigration DNA testing is now used by the Home Office to confirm disputed relationships, and GPS-ibd is used to process samples in immigration DNA tests that involve kinship disputes, for example, where an unaccompanied child refugee has applied to be reunited with family members already in the country. The software has also been used in immigration cases by various border and home security authorities in several other countries including Norway and the State of Kuwait [5.7].

Eurofins also uses the software to routinely carry out DNA identification of bodies for Coronial purposes. Coroners are tasked to not only determine the cause of death but to investigate or confirm the identity of unknown person(s) who has been found dead within their jurisdiction [6]

News media often focus on new DNA-based techniques that are helping authorities pursue investigations that defy conventional approaches, or even reopen investigations that were suspended long ago. Advances in the field of forensic DNA testing are helping to solve a broadening range of difficult cases, including unidentified persons, sexual assaults, and homicides. GPS-ibd software is included in these new techniques and have enabled forensic cold case reviews. A specific case from 2019, North Yorkshire Police confirmed the identity of a south-east Asian woman whose body was discovered in a remote location on the Pennine Way, between Pen-y-ghent and Horton in Ribblesdale, on 20 September 2004. GPS-ibd was used to assist with her identification. [5.8]

Over the course of Troubles in Northern Ireland there were 16 people who "disappeared". The Provisional IRA admitted to being involved in the forced disappearance of nine of the sixteen victims, mostly in a statement issued in 1999. One victim was admitted to by the INLA. No attribution has been given to the others. The Independent Commission for the Location of Victims' Remains (ICLVR) was established under The Good Friday Agreement between the UK and Irish Governments in connection with the affairs of Northern Ireland, in order to locate 16 missing Irish and British people presumed murdered during The Troubles. To date, the remains of twelve of 'the disappeared' have been recovered, ten of whom have been recovered through the ICLVR's efforts. The GPS-ibd software has been used extensively in this REF period to assist the ICLVR in identifying the 'disappeared' [5.7], [5.9].

Familial DNA analysis is not only helpful in the identification of the criminal, but also for the identification of the deceased in case of mass disaster victim's identification or in other cases, such as plane crash etc. In this REF period the software has been used in identifying bodies following disasters, perhaps the highest profile case was to assist in identifying the victims of the Grenfell Tower fire in 2017 which caused 72 deaths and was the worst UK residential fire since the Second World War [5.6]. The choice of GPS-ibd for this work was in part due to its success in the DNA identification of victims of the South-East Asia tsunami in Thailand and the Air France flight 447 air crash investigation in earlier times.

There are few more impactful roles for science/statistics than its application to crime. The impact on the public is direct and palpable. There are few projects (perhaps with the exception of medicine) where the public benefit is so obviously there.

5. Sources to corroborate the impact (indicative maximum of 10 references)

5.1 Vidavia Group Facebook (5 August 2015) 'we've managed to sign an agreement with City University to take forwards the brilliant FSS-ibd (aka FSS DNA Lineage) software'. Retrieved from: https://www.facebook.com/VIDAVIAgroup/posts/100547386966595

5.2 Vidavia Group – Portfolio. Retrieved from http://www.genepoolsystems.com/

5.3 Maguire, Chris N. (2015) Development Timeline for GPS-ibd, Internal document documenting product history for VIDAVIA Media. 16th October, (5 pages) available on request.

5.4 Email communication between the Director, VIDAVIA MEDIA S.L. and City, University of London Tech Transfer Office [Received Tue 24/04/2018 11:46] – available on request.

5.5 LGC Group Company News - LGC accredited for Relationship Analysis Services using ground breaking software. Published on 29 MAR 2016. Retrieved from: https://web.archive.org/web/20160701000000*/https://www.lgcgroup.com/about-us/media-room/latest-news/2016/lgc-accredited-for-relationship-analysis-services/

5.6 Personal testimony from Dr Tim Clayton MBE, Senior Reporting Scientist, DNA Science Lead, Eurofins Forensic Services [received 18 February 2020]

5.7 Cases that used GPS-ibd. Personal Communication and testimony from Director of ForGenetica Consultants Ltd. [21 January 2020] - available on request.

5.8 Woman identified in 2004 Pen-y-Ghent body cold-case investigation, North Yorkshire Police News [Last modified: 19 March 2019 at 02:22pm] Testimony from Dr Tim Clayton [5.6] confirms that GPS-ibd software was used to assist with her identification.

5.9 Henry McDonald, Remains confirmed as IRA 'disappeared' Séamus Wright and Kevin McKee, The Guardian, Tue 8 Sep 2015 13.49 BST. Testimony from Dr Maguire [5.7] states that GPS-ibd was used to forensically identify the bodies.

Note: LGC Forensics was acquired by Eurofins in 2017. Dr Clayton's testimonies above have different employer affiliations depending on the year in which they were given.