Institution: University of York		
Unit of Assessment: 10 - Mathematical Sciences		
Title of case study: Chemometric methods for food and environmental studies		
Period when the underpinning research was undertaken: 2003 - 2020		
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:**
Julie Wilson	Professor	Oct 1999 - present
Simon Poulding	Research Assistant	Sep 2006 – Jul 2011
Elizabeth Dickinson	Research Assistant	Apr 2016 – Jul 2019
Period when the claimed impact occurred: 2014 – 2020		
Is this case study continued from a case study submitted in 2014? N		

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

Mathematical techniques developed by Wilson's group at the University of York have supported projects with Fera Science Ltd worth over GBP30,000,000, maximising the useful information extracted from the analysis of complex mixtures of metabolites. Algorithms developed by Wilson for processing the mega-variate datasets obtained by analytical chemistry are now used routinely by Fera scientists. They have been applied in food safety and authentication studies, for the identification of drought and disease resistant crop varieties and to determine biomarkers for animal disease.

2. Underpinning research (indicative maximum 500 words)

Analytical chemistry methods such as Nuclear Magnetic Resonance (NMR) and Liquid Chromatography-Mass Spectrometry (LC-MS) allow the simultaneous identification of a wide range of small molecules, or metabolites, to provide characteristic ''fingerprints'' detailing the relative concentrations of compounds present in a sample. Each sample may produce thousands of data points, and so requires sophisticated data reduction techniques to extract useful information. Chemometrics applies mathematical methods using statistics and pattern recognition to these metabolomic fingerprints, enabling samples to be classified, anomalies recognized and markers for different biological states, for example diseased or droughted, to be identified. Various novel algorithms have been developed by Wilson's group in response to specific industrial challenges identified at Fera Science Ltd [R1] – [R6].

Changes in experimental parameters such as temperature, pH and ionic strength can result in unwanted NMR peak shifts, making inter-sample comparison impossible and necessitating more detailed specific modelling of spectral peaks. Using a non-decimated wavelet transform, Wilson's adaptive binning algorithm [R1] provides variable-length bins, which correspond directly to peaks in the spectra and thus facilitate interpretation. Noise regions are recognised and excluded, significantly reducing variation within a biological class (for example, disease state) in comparison to fixed-width binning.

A novel feature extraction method developed by Wilson's group now allows two-dimensional NMR techniques, such as Heteronuclear Single Quantum Coherence (HSQC) and Heteronuclear Multiple Bond Correlation (HMBC), to be used in the analysis of complex mixtures. The method uses a modified Lorentzian function to model peaks in ^{1}H-^{13}C HSQC spectra [R2]. This provides elliptical footprints corresponding to peaks in the spectra and allows the deconvolution of overlapping peaks, giving a dramatically reduced set of variables suitable for metabolomic analyses. With phase-cycled HSQC, systematic noise needs to be removed before feature extraction. Despite its superior sensitivity, this technique has previously been limited by the presence of noise ridges, which can mask genuine peaks for low-concentration compounds. The group's Correlated Trace Denoising (CTD) algorithm [R3] takes advantage of the systematic nature of this so-called t1 noise and, unlike other methods for t1 noise removal that have specific pre-requisites, CTD can be used regardless of complexity and the number of peaks in a spectrum.

To allow necessary calibrations and cleaning of the instrument, LC-MS spectra are often acquired batch-wise, which introduces further sources of variation. Quality control (QC) samples are frequently employed to assess and correct for this variation but the non-linearity of the response can result in substantial differences between the recorded intensities of the QCs and experimental samples. This makes the required adjustment difficult to predict and can even exacerbate the problem by introducing artificial differences. Wilson's correction method identifies time series trends using all samples and does not rely on the availability of suitable QC samples [R4]. The method has been shown to reduce differences between replicate samples and thereby highlight differences between experimental groups previously hidden by instrumental variation.

The software package MetaboClust [R5] was developed by Wilson's group to provide an interactive approach to metabolomic time-course analyses and can be used to apply data correction techniques, generate time-profiles, perform exploratory statistical analysis and assign tentative metabolite identifications in a workflow with visual feedback at all stages of analysis. Clustering can be used to group metabolites in an unbiased manner, allowing pathway analysis to score metabolic pathways, based on their overlap with clusters showing interesting trends. Trends in time series can reveal differences between groups, for example healthy versus diseased, that analysis of individual observations (a metabolic snapshot) cannot. Clustering methods developed for transcriptomics have been integrated into a pipeline for the analysis of metabolomic time series and used to identify markers for stress in plants [R6].

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

[R1] *Davis, R.A., Charlton, A.J., Godward, J., Jones, S.A., Harrison, M. and Wilson, J.C., 2007. Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform. *Chemometrics and intelligent laboratory systems*, *85*(1), pp.144-154. DOI:10.10.16/j.chemolab.2006.08.014

[R2] *McKenzie, J.S., Charlton, A.J., Donarski, J.A., MacNicoll, A.D. and Wilson, J.C., 2010. Peak fitting in 2D 1 H–13 C HSQC NMR spectra for metabolomic studies. *Metabolomics*, *6*(4), pp.574-582.DOI: 10.1007/s11306-010-0226-7

[R3] *Poulding, S., Charlton, A.J., Donarski, J. and Wilson, J.C., 2007. Removal of t1 noise from metabolomic 2D 1H–13C HSQC NMR spectra by correlated trace denoising. *Journal of Magnetic Resonance*, *189*(2), pp.190-199. DOI:10.1016/j.jmr.2007.09.004

[R4] *Rusilowicz, M., Dickinson, M., Charlton, A., O'Keefe, S. and Wilson, J., 2016. A batch correction method for liquid chromatography–mass spectrometry data that does not depend on quality control samples. *Metabolomics*, *12*(3), p.56. DOI:10.1007/s11306-016-0972-2

[R5] *Rusilowicz, M.J., Dickinson, M., Charlton, A.J., O'Keefe, S. and Wilson, J., 2018. MetaboClust: Using interactive time-series cluster analysis to relate metabolomic data with perturbed pathways. *PloS one*, *13*(10), p.e0205968. DOI:10.1371/journal.pone.0205968

[R6] *Dickinson, E., Rusilowicz, M.J., Dickinson, M., Charlton, A.J., Bechtold, U., Mullineaux, P.M. and Wilson, J., 2018. Integrating transcriptomic techniques and k-means clustering in metabolomics to identify markers of abiotic and biotic stress in Medicago truncatula. *Metabolomics*, *14*(10), p.126. DOI:10.1007/s11306-018-1424-y

*= peer reviewed publication

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

The research described was carried out at the University of York in response to complex data analysis problems encountered by Fera Science Ltd, a science-based organisation which works across the agri-food supply chain [E1] and has over 7,500 government and commercial customers and provides services to customers in over 100 countries. Wilson has developed methodology which is now integral to Fera's work, and in routine use to provide more accurate

analyses. Fera have sponsored three PhD studentships and, according to the Head of Chemical and Biochemical Profiling at Fera, the collaboration has "provided substantial outputs, related largely to the translation of the outcomes from large-scale projects, in particular for Defra, the Food Standards Agency, the European Commission and BBSRC, into a commercial setting." He states that "The novel techniques developed as part of the collaboration underpin a Fera team delivering a range of commercial projects, which have an approximate value greater than GBP30M" and have "also been instrumental in supporting the international commercial sector" [E2].

Fera Science engage with many EU and Defra projects, combining policy and regulatory know-how with extensive testing and analysis capabilities [E1]. One project underpinned by the methods developed by Wilson's group is the GBP4,000,000 EU-funded ABSTRESS consortium co-ordinated by Fera Science. This consortium of 13 industrial and academic partners from 7 EU countries identified processes in plant biochemistry associated with the way drought and disease combine and exacerbate plant stress. The team produced new seed varieties for commercial breeding programmes "to breed a new generation of crops more able to cope with the challenges of climate change" [E3].

The Operations Manager for Chemical Contaminants and Food Integrity at Fera Science states that Wilson's batch correction method has "allowed us and our collaborators to assess these data sets more accurately and efficiently" in an EU-funded HEALS (Health and Environment-wide Associations based on Large Population Surveys) project, where the large number of samples required the data to be acquired in batches. This EUR12,000,000 project aimed to assess individual exposure to environmental pollution and predict health outcomes [E4]. The project involved a series of population studies across Europe including twin cohorts, tackling different levels of environmental exposure, age, windows of exposure and socio-economic and genetic variability. The research, covering ten EU and one African state, provided scientific advice on development of the protocols and guidelines needed to set up a larger European environment and health examination survey. The Operations Manager states that "The algorithms developed by Professor Wilson's group have allowed us and our collaborators to assess these data sets more accurately and efficiently" and that "Outputs from the HEALS project include associations with an individual's internal and external exposome and the onset of respiratory or motor neurological disease" [E4]. Regarding other Fera projects, he says that "batch correction has helped move the technology readiness level (TRL) from 1-3 to 4-6 in studies to identify potential metabolic markers within fish to help confirm origin and type of capture for the UK's Marine Management Organisation" [E4]. That is, the studies have moved on from 'experimental proof of concept' to 'technology validated and demonstrated in an industrially relevant environment'. This algorithm, along with others developed by Wilson's group, has recently been incorporated into Fera's Matlab-based analysis software, Metabolab. This now allows analytical chemists at Fera without any programming experience to implement the methods.

Together, the University of York and Fera Science Ltd are playing an important role in the international honey sector in relation to data analysis and food fraud. The Head of Chemical and Biochemical Profiling at Fera states that "This has led to high level representation in the UK and New Zealand Parliaments" [E2]. Biomarkers chosen to authenticate Manuka honey by the Ministry for Primary Industries (MPI) in New Zealand are disputed by honey producers from the Unique Mañuka Factor Honey Association (UMFHA) due to their potential instability during storage. Wilson's methods were used to demonstrate the problem with existing data and, following a teleconference involving representatives from both MPI and UMFHA, a protocol for a large-scale stability study with chemical analysis performed at Fera Science Ltd has been agreed [E5].

Honey, a product in high demand and short supply, has become a target for economically motivated food fraud. This may involve, for example, the addition of cheap sugar syrups to increase the volume of lucrative premium honey by claiming a false geographic origin or floral source. Nuclear Magnetic Resonance (NMR) has recently been used to detect the apparent

adulteration of honey with sugar syrups with highly controversial results [E6]. In collaboration with the UK Honey Association, Fera Science Ltd are now curating an extensive database to demonstrate the natural variance in honey from countries across the globe and provide reliable honey authenticity testing. The Head of Chemical and Biochemical Profiling at Fera says "The analysis of data in the database is entirely based on algorithms developed with Prof. Wilson. The approach and initial findings have been presented to the European Commission to help harmonise fraud detection approaches within the European honey sector" [E2].

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

[E1] https://www.fera.co.uk/our-science/active-r-and-d/eu-and-defra-projects (accessed 7/11/2020).

[E2] E-mail provided by Dr Adrian Charlton, Head of Chemical and Biochemical Profiling at Fera Science Ltd.

[E3] https://cordis.europa.eu/article/id/169571-strengthening-legume-crops (accessed 10/08/2020).

[E4] Letter provided by Dr Michael Dickinson, Operations Manager, Chemical Contaminants and Food Integrity at Fera Science Ltd.

[E5] E-mails regarding stability trials for Manuka honey biomarkers, involving the Ministry for Primary Industries (MPI) in New Zealand and the honey producers from the Unique Mañuka Factor Honey Association.

[E6] BBC news item showing controversial results obtained by NMR testing of Tesco honey. https://www.bbc.co.uk/news/uk-50551385 (accessed 6/11/20)