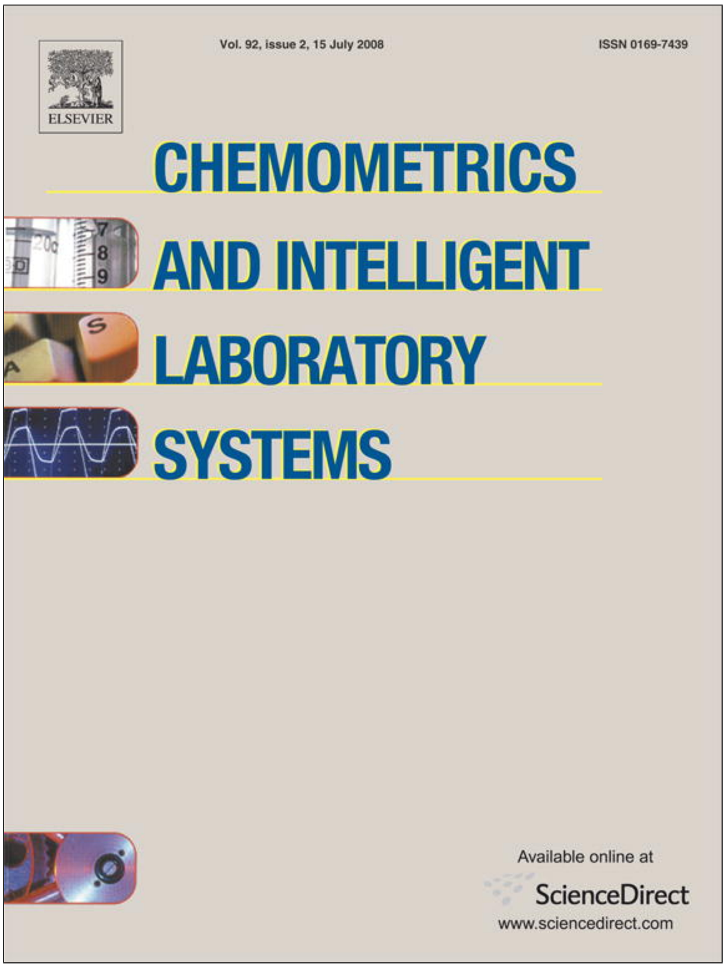

Vol. 92, issue 2, 15 July 2008

ISSN 0169-7439

ELSEVIER

CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS

Available online at
ScienceDirect
www.sciencedirect.com

Available online at www.sciencedirect.com

ScienceDirect

Chemometrics and Intelligent Laboratory Systems 92 (2008) 101 – 109

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Generalized multiplicative analysis of variance of kill kinetics data of antibacterial agents

Diako Ebrahimi [a],*, M. Sharif Sharifi [b], Stuart L. Hazell [b], D. Brynn Hibbert [a]

[a] *School of Chemistry, The University of New South Wales, Sydney, NSW 2052, Australia*
[b] *School of Biomedical and Health Sciences, University of Western Sydney, Campbelltown, NSW 2560, Australia*

Received 27 August 2007; received in revised form 12 January 2008; accepted 13 January 2008
Available online 18 January 2008

Abstract

Estimation of a large number of parameters to model higher order interaction terms limits the interpretability and therefore applicability of classic ANOVA models. Multiplicative models have been proposed to tackle this problem in data generated mainly by interactions. In this work a GEneralized Multiplicative ANalysis Of VAriance (GEMANOVA) method is applied to assess the bactericidal activity of novel antimicrobial agents isolated from plant extracts in different structure and oxidation forms and different concentrations on three genera of bacteria. While the applicability of ANOVA is restricted due to the complex interaction among the factors, GEMANOVA is shown to return robust and easily interpretable models which conform to the actual structure of the data. This study is the first application of GEMANOVA to model the data from the field of microbiology and the first GEMANOVA model in which more than two multi-way terms are used and interpreted.

Keywords: Generalized multiplicative analysis of variance; Interaction; Multi-way analysis; Antibacterial; Kill kinetics

1. Introduction

In classic ANalysis Of VAriance (ANOVA) data are modeled using main effects of each factor and as few as possible low order interactions [1]. The presence of significant high order interaction terms can result in ANOVA models which are difficult to interpret. Use of so-called “multiplicative” terms (described in Section 2 below) has been shown to simplify the model and thus its interpretability [2]. ANOVA of two-factor experiments in which multiplicative terms are used have been described by different authors [3–5]. In these works, the data were modeled using the main effects and interactions. Decomposition of interactions using principal component analysis (PCA) then resulted in models with greater interpretability. The idea of modeling the interactions by a linear decomposition has been extended to more than two-way interaction decomposition [6]. A general framework of these methods known as “GEMANOVA, GEneralized Multiplicative ANalysis Of VAriance” was first proposed by Bro and describes modeling of experimentally designed data such as those designed with full or fractional factorial designs [7,8]. Decomposition of multi-way data used the PARAFAC (parallel factor analysis) algorithm, which results in unique models without rotational ambiguity [2]. In the first applications of GEMANOVA [8,9] the enzymatic browning caused by polyphenol oxidase was monitored as a function of pH, temperature, substrate and amounts of oxygen and carbon dioxide. Experimentally designed data sets (multi-way arrays) were decomposed using GEMANOVA to interpret the effects of each factor. Later Bro proposed a new algorithm that allowed both main effects and multiplicative interaction terms of any order in the model [10]. An example from the beef industry was presented in that work. In a similar study GEMANOVA was used to model the effect of different packaging and storage conditions on the color stability of cured, cooked ham [11]. Here we report the first application of GEMANOVA in which

* Corresponding author. Tel.: +61 2 9385 6621; fax: +61 2 9385 6141.
E-mail address: diako@unsw.edu.au (D. Ebrahimi).

0169-7439/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.01.001

more than one multi-way term is required to model the data adequately.

"Kill kinetics" studies [12] are used to determine the killing rate of bacteria by antimicrobial agents. In kill kinetics bacteria are incubated with an antibacterial agent and the number of viable bacteria is traditionally plotted against time. A more effective antibacterial agent gives a steeper decline in the graph and thus smaller area under the curve.

The aim of the present work is to compare the GEMANOVA model with classic ANOVA in interpretation of the kill kinetics data of a microbiological system with complex interactions among its factors. In this context the activity of an already known bactericidal (*β-myrcene*) and active ingredients of four natural gums (coded as *B*, *R*, *S* and *T* due to intellectual property restrictions) in two structures (monomer and polymer), two oxidation forms (oxidized and non-oxidized) and two concentrations (1 and 5 × MIC, Minimum Inhibitory Concentration) [13] were investigated on two Gram-negative (*Escherichia coli* (*E. coli*) type 1 UNSW 048200 and *Helicobacter pylori* (*H. pylori*) strain 26695) and one Gram-positive *Staphylococcus aureus* (*S. aureus*) UNSW 056201) bacteria.

2. Methodology

A two-factor ANOVA may be formulated as [10]

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}; \quad i = 1,\ldots,I \quad j = 1,\ldots,J \qquad (1)$$

where x_{ij} and e_{ij} are the observation and model residual at *i*th and *j*th instances of the first and second factors respectively, μ is the population mean, a_i and b_j are the effects of two different factors with $i=1$ to I and $j=1$ to J instances respectively. $(ab)_{ij}$ is the interaction (joint effect) of *i*th and *j*th instances of the first and second factors respectively. To investigate the main effect of a factor with I instances, I parameters are estimated in a classic ANOVA model. Regarding the interaction effects, $I \times J$ parameters are calculated in order to include a two-way interaction term in the model. The number of parameters dramatically increases for higher order interaction terms which therefore can be difficult to interpret. Consequently, a model using the main effects and a few lower order interactions, if necessary, is sought.

Second-order (two-way) interaction terms can be simplified by decomposing the array of data using PCA to build a multiplicative model [3–5]. The complexity of the model is dramatically reduced by computing fixed loadings for each factor. A multiplicative model of a two-way interaction requires $I+J$ parameters as shown in Eq. (2) for a bi-linear model with two loadings (called profiles), a_i and b_j.

$$x_{ij} = a_i b_j + e_{ij}; \quad i =,\ldots,I \quad j = 1,\ldots,J. \qquad (2)$$

The idea of modeling a second-order interaction term using a two-way model such as PCA has been extended to modeling higher order interactions using multi-way models. These models are particularly useful to study experimentally designed data with more than two factors as shown in Fig. 1a with three factors F1, F2 and F3 studied at *I*, *J* and *K* instances. A full factorial design in which the experiments are performed in all possible combinations of instances of factors forms a cube of data. Each factor is presented by one mode with *I*, *J* or *K* elements corresponding to the number of instances of each factor. Decomposition of this array by a multi-linear model such as PARAFAC returns a number of components each consisting of three loadings vectors. Each loadings vector represents one of the three factors F1, F2 or F3. In this example number of components is four, thus the data is modeled by four third-order multiplicative interaction terms.

The multiplicative models obtained by a PARAFAC algorithm to describe the variance of data using higher order interaction terms are referred to as GEMANOVA [7,8,10]. A drawback of using the original GEMANOVA model was that the data was modeled exclusively by multi-way terms comprising all the factors. It means that to study three factors only third-order interactions were used, to study four factors only fourth-order interaction were used and so on. Main effects and lower order interaction terms could not be included in the original PARAFAC/GEMANOVA algorithm. Bro has proposed a new version of GEMANOVA which takes advantage of both

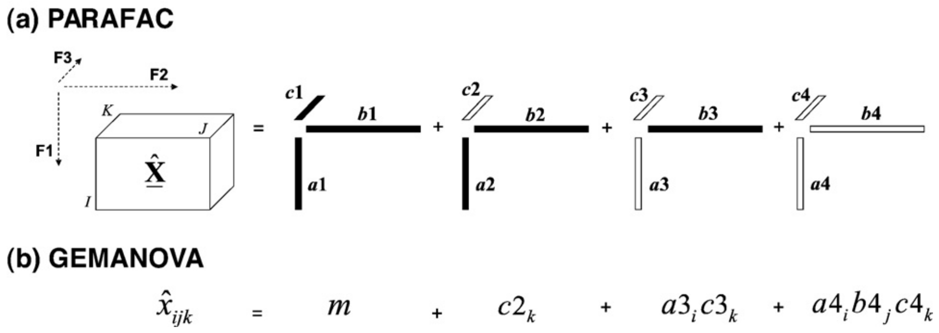

Fig. 1. Schematic presentation of the data in a full factorial design experiment with three factors F1, F2 and F3 with *I*, *J* and *K* instances using a multi-way array which can be decomposed by PARAFAC. In this example the variance of the data is modeled by four three-way interaction terms. This figure also presents a GEMANOVA model with four terms using a constrained three-way PARAFAC model with four trilinear components. $\hat{\underline{X}}$ is the estimated array using a PARAFAC model. a_1–a_4, b_1–b_4 and c_1–c_4 are loadings vectors of PARAFAC model. Unity loadings are shown by filled vectors and non-unity loadings by open vectors in Fig. 1a. $\hat{x}_{ijk}$ is the estimated response at the *i*th instance of factor F1, *j*th instance of factor F2 and *k*th instance of factor F3 using a GEMANOVA model (Fig. 1b) with a constant term (m), a main effect of F3 ($c2_k$), a second-order interaction between F1 and F3 ($a3_i c3_k$) and a third-order interaction among F1, F2 and F3 ($a4_i b4_j c4_k$).

multi-linear and ANOVA models. In his approach the data can be modeled using any combinations of main effects and lower to higher order interaction terms. The general structure of this GEMANOVA model for two factors is given by

$$x_{ij} = \mu + a_i + b_j + \sum_{r=1}^{R} c_{ir} d_{jr} + e_{ij} \tag{3}$$

where a_i and b_j are the main effects identical to those in the ANOVA model of Eq. (1). The next term shows the two-way (second-order) interaction as a summation over R components (it is worth noting that the symbol R used here is different from the code R used earlier for one of the instances of factor 'substance'). Note that except for once, where a GEMANOVA model with a four-way term plus a main effect was reported [10], in the GEMANOVA models used in literature [8,9,11] so far one multi-way term ($R=1$) has been adequate to model the data. However several multi-way terms of arbitrary dimensions are allowed using the new GEMANOVA algorithm. This means for N factors a full model would include the main effects and all combinations of two-to N-order interaction terms. Note that similar to ANOVA only significant terms are to be included in any GEMANOVA model. A constrained PARAFAC algorithm is used to build the GEMANOVA model in the new version. Each term in the GEMANOVA model can be seen as a multi-way component of a PARAFAC model in which some (or none) of the profiles are constrained. For example in a data set with three factors F1, F2 and F3, if a two-way interaction term between F1 and F3 is required in the model, a three-way PARAFAC component in which the values of the F2 profile is set to unity can be used. To clarify this, a GEMANOVA model with four terms (a constant, main effect of F3, two-way interaction between F1 and F3 and three-way interaction among F1, F2 and F3) is graphically displayed in conjunction with its corresponding constrained PARAFAC model in Fig. 1. As displayed, each term in GEMANOVA corresponds to one multi-way component in PARAFAC. The first GEMANOVA term in Fig. 1b is a grand level (a constant) which is obtained using a three-way PARAFAC component with three unity loadings vectors (Fig. 1a). The second term which is the main effect of F3 is given by a three-way PARAFAC component in which the loadings vectors of F1 and F2 are unity. The third term is a two-way interaction between F1 and F3 described by a PARAFAC component with only one unity loadings vector, F2. Finally the last term, a third-order interaction is modeled by a normal three-way PARAFAC component without constraints.

3. Experimental

The experimental section of this work will be published as a patent and therefore receives intellectual property restrictions. For this reason only a brief introduction of few experimental details not bounded by intellectual property restrictions will be provided in the present paper. In addition for the new chemical extracts displaying antibacterial activities coded names are used rather than their actual chemical names. Further information regarding the chemical methods used and the structure and characteristics of antibacterial agents are found in the original reference [14].

3.1. Experimental design

A full factorial design was performed in triplicate to study kill kinetics of five bactericides (*β-myrcene*, natural gums *B*, *R*, *S* and *T*) in two structures (monomer and polymer), two oxidation forms (oxidized and non-oxidized) and two concentrations (1 and 5 × MIC, Minimum Inhibitory Concentration) on three bacteria (*S. aureus*, *E. coli* and *H. pylori*). Thus $3 \times 5 \times 2 \times 2 \times 2 \times 3 = 360$ experiments were performed.

3.2. Preparation of polymers

β-myrcene (237 g), cyclohexane (385 g), and sec-butyl lithium (4 ml, 1.3 M in cyclohexane) were maintained at 60 °C for 1 to 4 h to obtain polymyrcene which was then precipitated with methanol and dried under vacuum to constant weight [15]. The polymeric fractions of the natural gums were isolated from the resin by dissolving the resins with dichloromethane [16] or anhydrous diethyl ether. The collected polymer was then analysed by NMR to confirm the structure and by gel permeation chromatography (GPC) to determine their molecular weight [14].

3.3. Oxidation

A sample of the polymer fraction (50 mg) was dissolved in dichloromethane (2 ml) and tested for the presence of carbonyl groups by adding a few drops of the solution to 1 ml of 2,4-dinitrophenol (2,4-DNP) reagent. Finely powdered polymer (1.00 g) was suspended in 50 ml of distilled water and analytical grade air bubbled through the suspension for 24 h. The solid was filtered off and allowed to air dry (1.11 g). The aerated, dried polymer (50 mg) was then dissolved in dichloromethane (2 ml) and oxidation confirmed by adding a few drops of the solution to 1 ml of 2,4-DNP reagent. The same procedure was carried out to oxidize synthetic polymyrcene and monomer fractions [14].

3.4. Minimum inhibitory concentration (MIC)

The MIC values were determined against *H. pylori* strains 26695, *E. coli* and *S. aureus* using a broth micro-dilution method [14].

3.5. Kill kinetics

To obtain the kill kinetics data samples of each bacterium were incubated separately with five antimicrobial substances in all combinations of structure, oxidation form and concentration (MIC) of substances. Bacteria were then counted in each experiment every hour for 48 h [14]. For *H. pylori* kill kinetics was performed with static liquid cultures containing Isosensitest broth (Oxoid) supplemented with 5% horse serum (Oxoid). The inoculum was harvested with Isosensitest broth from 36 h cultures grown on Campylobacter Selective Agar (CSA) [14]. In *E. coli* and *S. aureus* cases a 100 ml Isosensitest broth (Oxoid) culture was inoculated with a 10% inoculum from *E. coli* or *S. aureus* culture. The cultures were allowed to grow to stationary phase which was determined by recording the Optical Density 600 (OD600) of the culture. The inoculums were then adjusted with the culture medium to give a starting concentration of 1×10^8 M [14]. Each culture was incubated for 2 h to allow recovery of the bacteria before the test compounds were added at their respective 1 MIC and 5 MIC concentrations. Control cultures at 1 and 5 MIC containing appropriate solvent/s were also performed.

3.6. Data analysis

The area under a kill kinetics curve describes the biological activity of a substance in terms of both the extent and speed of killing bacteria. A substance which kills more bacteria in a shorter time is more effective and results in a lower kill kinetics area. In this work the five factors studied will be termed 'substance', 'bacterium', 'structure', 'oxidation form' and 'concentration'. Areas under these kinetic curves recorded for 48 h were calculated and arranged as vectors for analysis by ANOVA and as five-way arrays of substance × bacterium × structure × oxidation form × concentration for analysis by GEMANOVA. Three ANOVA models were built using replicates 1 and 2, 1 and 3 and 2 and 3 respectively. This allowed first, the inclusion of the main effects and all two-to five-order interaction terms in the models and secondly, the assessment of the prediction ability of the models using the replicate which was not used in each model as test data. The models were refined by removing insignificant terms at the 95% probability level (*F*-test of the mean square due to the term against the residual variance) and a new ANOVA model was built with the remaining terms in each case. These optimized ANOVA models were then compared with the optimum GEMANOVA model obtained as explained below.

Several GEMANOVA models including two-to five-way multiplicative and constant terms were performed on the five-way data array to select the best model in terms of RMSEC (root mean square error of calibration), RMSECV (root mean square error of cross validation) and correlation coefficient between the response and cross validated estimated response (r^2CV) [2]. While the residual sum of squares of calibration describes the ability of the models to fit the data, it is also possible to compare the predictive ability of models by cross validation. In a leave-one-out cross validation one sample is left out and a calibration model with the rest of samples is built. The excluded sample is then predicted using the model. This process continues until all the samples are left out and predicted. The square root of the average residual sum of squares and cross validated residual sum of squares are calculated and termed RMSEC and RMSECV respectively. They are given by

$$\mathrm{RMSEC}(V) = \sqrt{\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M}\left(\hat{y}_{ijklm} - y_{ijklm}\right)^2}{N}} \tag{4}$$

where N is the number of samples in the calibration set (it is worth noting that the letter N used here should not be confused with that used for the number of factors in an ANOVA model). y_{ijklm} is the response at *i*th, *j*th, *k*th, *l*th and *m*th instances of factors 'substance', 'bacterium', 'structure', 'oxidation form' and 'concentration', respectively. I, J, K, L and M are the number of instances of these factors respectively. In RMSEC $\hat{y}_{ijklm}$ are the values of the predicted responses when all samples are included in the model and in RMSECV, $\hat{y}_{ijklm}$ are the prediction of samples not included in the calibration.

The optimum GEMANOVA model was compared with the optimum ANOVA models using four figures of merit: r^2 (correlation coefficient between the estimated response and the measured response of the replicate not included in the model), the number of estimated parameters, and the values of RMSEC and RMSEP (root mean square error of prediction, given by Eq. (4) except that y_{ijklms} is the response of replicate which is not included in the analysis).

All mathematical manipulation was performed on a personal computer (Pentium 4, 3.2 GHz) running Windows (Microsoft Inc) and Windows Office 2002 (Microsoft Inc, USA). All calculations were implemented in MATLAB®, Version 7.1 (The MathWorks Inc., USA). GEMANOVA MATLAB code was obtained from R. Bro at http://www.models.kvl.dk/source. The GEMANOVA software was validated using the standard data set of "Colour_of_Beef.mat" obtained from http://www.models.kvl.dk/research/data. The results were verified against published data.

4. Results and discussion

4.1. ANOVA

As described in the Methodology section, the number of ANOVA parameters for the main effect of each factor equals the number of instances of that factor. For example factor 'substance' has five instances, thus five parameters are estimated for its main effect. When interaction terms are concerned the number of estimated parameters is the product of the number of instances of interacting factors. For example to model a second-order interaction between factors 'substance' and 'bacteria' with 5 and 3 instances in each, 15 (=5×3) parameters are estimated. In this work, to model all the main effects and two-to five-order interaction terms, 647 parameters were estimated.

In each set of experiments, to cover all the combinations of instances of factors, 120 experiments were performed. To provide sufficient degrees of freedom to estimate the parameters, ANOVA models were fitted to pairs of replicate data sets. (In contrast to ANOVA models a single replicate has sufficient data for the GEMANOVA analysis.) Results of the ANOVA models of replicates 1 and 2, 1 and 3 and 2 and 3 showed that the main effect of 'structure' and the two-way interaction between 'structure' and 'concentration' were not significant (at 95% level) for any of the models. The remaining terms (29 out of total 31) were significant in at least one model. The presence of many significant higher order interactions in

Table 1
Figures of merit (number of parameters, RMSEC, RMSEP and r^2) of ANOVA and GEMANOVA models averaged on the results of three analyses in each

Model	No. of parameters[a]	RMSEC	RMSEP	r^2
Classic ANOVA	442, 566, 540	480	801	0.93
GEMANOVA	23, 23, 23	912	1333	0.93

[a] Number of parameters for three ANOVA models of pairs 1 and 2, 1 and 3 and 2 and 3 of replicates and three GEMANOVA models of replicates 1, 2 and 3 are shown in order.

the classic ANOVA models required the estimation of a large number of parameters (more than 400, see Table 1) which restricted the interpretability and so applicability of ANOVA in this case. In addition, the three ANOVA models showed different statistical significances for ten (out of 31) of the terms. For example while the second-order interaction between 'substance' and 'concentration' was returned non-significant by ANOVA of replicates 1 and 2 and 2 and 3, it was significant using replicates 1 and 3. These contradictions using ANOVA models of different replicates might suggest that classic ANOVA is not robust for this data set.

4.2. GEMANOVA

Several GEMANOVA models including two to five-way multiplicative terms and constant terms were applied to the first replicate data set. For each model the RMSEC and RMSECV were calculated (Fig. 2). Each model in the figure is referenced by its estimated terms using the default format of the GEMANOVA routine in which an *n*-digit binary number codes an *n*-way term in the GEMANOVA model. Each digit represents one of the modes for which 1 indicates that the loadings vector of the mode is unity (and therefore the mode is absent in the term) and 0 means the mode is present in the term. As an example model 00010 is a GEMANOVA model with a single term which is fourth-order (four-way) interaction between modes one (first digit), two (second digit), three (third digit) and five (fifth digit), that is, modes 'substance', 'bacterium', 'structure' and 'concentration'. 'oxidation form' mode is not present here therefore is given 1 (fourth digit). Similarly model 00000-01111 displays a model with two multi-way terms. The first term is a five-way interaction between all five factors (all zeros) and the second term is the main effect of substance (zero in the first position).

As displayed in Fig. 2 both RMSEC and RMSECV values of GEMANOVA models with two and three multi-way terms are significantly lower than those with only one multi-way term including model 00000. This implies that more than one multi-way term is needed to model this data set. The RMSEC value of the GEMANOVA model with three five-way terms (model 00000-00000-00000) is less than those of models with two multi-way terms, but its RMSECV value is greater than the RMSECV values of some of the models with two terms. This implies that a better fit (smaller RMSEC) obtained by inclusion of more terms does not necessarily guarantee a better prediction (smaller RMSECV). Consequently a GEMANOVA model with two multi-way terms is an optimum model for this data set. Among the models with two multi-way terms, those which have at least one five-way term display lower model errors (RMSEC and RMSECV) when compared with those with lower order interaction (two-, three- and four-way interaction) terms and main effects (models 00011-11100 to 10000-01111). To select among the GEMANOVA models with two multi-way terms of which at least one is five-way, the r^2CV of these models were calculated and compared (see the right hand side of Fig. 2). As

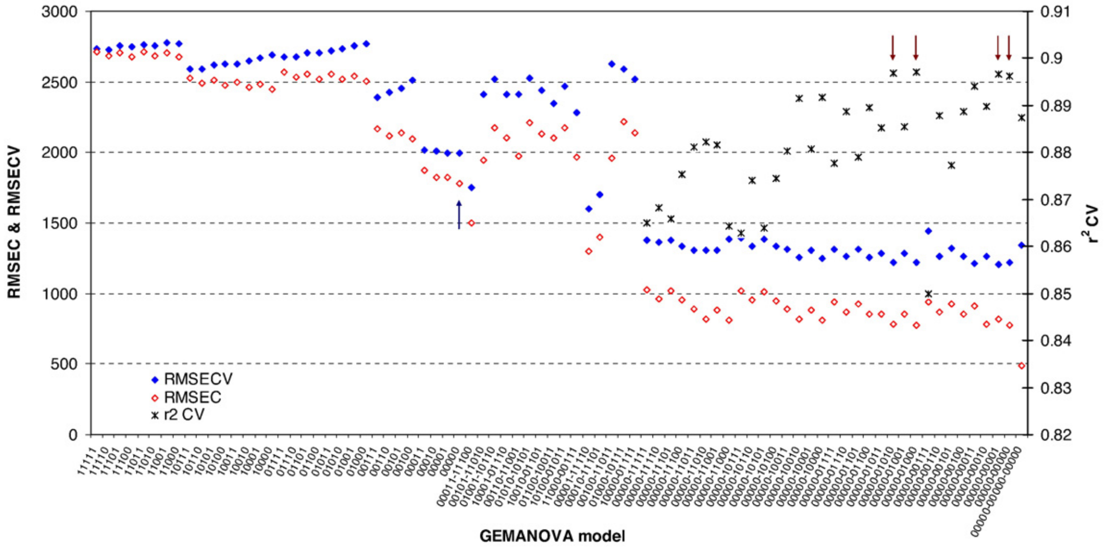

Fig. 2. RMSEC, RMSECV and r^2CV results of different GEMANOVA models using the data of the area under the kill kinetics curves of replicate one. Data is of size 5×3×2×2×2. The result of a one-component five-way GEMANOVA model (model 00000) is shown with an arrow pointing up. The best four models in terms of RMSEC, RMSECV and r^2CV are shown by arrows pointing down.

indicated, four models of 00000-01010, 00000-01000, 00000-00001 and 00000-00000 show the lowest errors (RMSEC and RMSECV) and the highest r^2CV, among which 00000-01010 is the most parsimonious model. While the number of parameters in the models 00000-00000, 0000-00001 and 00000-01000 are 28, 26 and 26 respectively, 23 parameters are estimated in the model 00000-01010 (5+2+2+2+3+5+0+2+0+2). Because model 00000-01010, provides the same quality in terms of RMSEC, RMSECV and r^2CV with fewer parameters, it is preferred over the other models. This model comprises one five-way interaction term including all the factors of substance, bacterium, structure, oxidation form and concentration and one three-way interaction term including substance, structure and concentration. The three-way term can be regarded as a five-way term in which bacterium and oxidation form have no effect and therefore the parameters of these terms are constant (=1). The effects of each factor (a mode of the five-way array) in this model are separated and given in separate loadings plots (Fig. 3) representing the effects of 'substance', 'bacterium', 'structure', 'oxidation form' and 'concentration'.

As may be seen in the GEMANOVA loadings (parameters) plots, substances exhibit different kill kinetics behavior. Differences between substances are more evident in the five-way term than in the three-way term. Substances are grouped into two categories by the five-way term: *β-myrcene* (Bm) and *S* show a low bactericidal effect (large loadings which are proportional to the area under the kill kinetics curve) while *B*, *R* and *T* exhibit greater antibacterial activity. On the other hand only minor differences between substances are observed from inspection of the three-way term. In the bacterium loadings plot for the five-way term, *S. aureus* differs from *E. coli* and *H. pylori*. The modes 'bacterium' and 'oxidation form' do not contribute to the three-way term, so using this term no differences between the bacteria or oxidation forms can be studied. Regarding the structure term two opposite effects are seen in the loadings plots. While the five-way term shows that polymerization reduces the bactericidal effect of substances, polymer forms of the substances are shown by the three-way term to be more effective. The five-way term shows that substances kill bacteria more effectively when they are in the oxidized form. As expected, greater activity is shown by the higher concentrations (5 MIC) of substances using both five- and three-way terms.

The estimated response (area under the kill kinetics curve) at a particular instance of each factor equals the sum of the products of the GEMANOVA parameters. As an example, the estimated response due to the incubation of *H. pylori* with the substance *B* in its polymer and oxidized form and at 5 MIC is given by

$$\begin{aligned}&\left(0.001\times-0.059\times0.999\times0.583\times1.837\times10^4\right)\\&\quad+\left(0.512\times1.000\times0.511\times1.000\times0.550\times10^4\right)\\&\qquad=-0.631+1438.976=1438.345.\end{aligned}$$

The first and second brackets consist of the coordinates in the five- and three-way terms respectively (see Fig. 3). Interestingly, the response is mainly estimated by the three-way term in this example (|−0.631|≪|1438.976|). The reason is that the value of the GEMANOVA parameter for *B* in the five-way term is very small (0.001). In fact the values of the parameters for *R* and *T* in the five-way term are also very close to zero. Therefore the kill kinetics data of these three antibacterial agents is explained almost exclusively by the three-way term. Regarding the bacterium loadings plot, the values of the GEMANOVA parameters for *E. coli* and *H. pylori* in the five-way term are very small; thus the effects of these two bacteria are described mainly by the three-way term. The same argument applies to the effect of structure in which the responses of the monomer shape with all combinations of the instances of other factors are described solely using the three-way term. As a result of domination of the responses by the three-way term for 'substance' = *B*, *R* and *T*, 'bacterium' = *E. coli* and *H. pylori* and 'structure' = monomer, the following conclusions are reached:

(a) Substances *B*, *R* and *T* kill all three species of bacteria better when they are in the polymer shape and at a concentration of 5×MIC. The oxidation form does not have an effect on their activity.

(b) Bacteria *E. coli* and *H. pylori* are more vulnerable when each substance is in the polymer form and at the higher concentration. *R* and *β-myrcene* (Bm) are the most and least effective substances respectively for these two bacteria. Sensitivity is not affected by the oxidation form of the antibacterial substances.

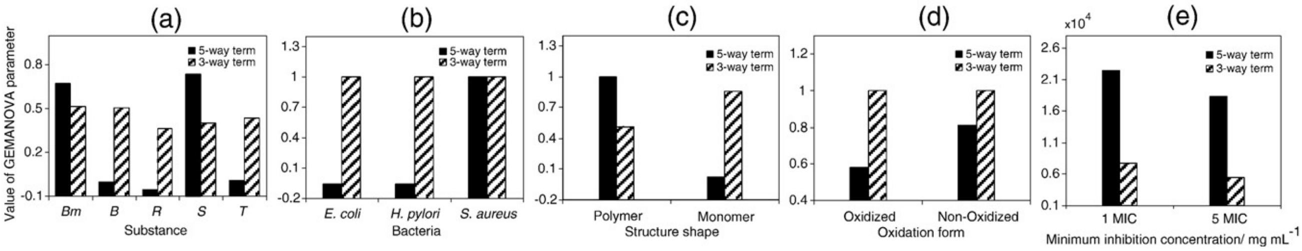

Fig. 3. Parameters of multiplicative terms from a GEMANOVA model with a five-way interaction of substance, bacteria, structure, oxidation form and concentration and a three-way interaction of substance, structure and concentration. Effects of (a) substance, (b) bacteria, (c) structure, (d) oxidation form and (e) concentration are calculated using a data of size (5×3×2×2×2) from replicate one.

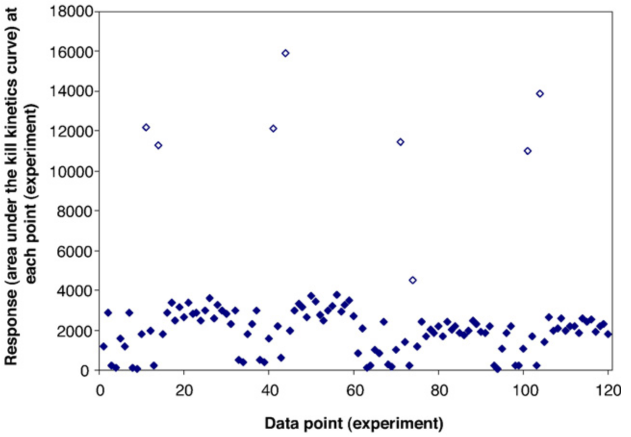

Fig. 4. Area under the kill kinetics curves of 120 experiments. Each point is an experiment performed in different combinations of instances of five factors of substance, bacteria, structure, oxidation form and concentration. Open diamonds show the data points which are modeled mainly by a five-way term and closed diamonds display the data points which are modeled mainly by a three-way term.

(c) The monomer shapes of all substances exhibit the same activity on the three studied bacteria. Oxidation form does not have an impact on the effectiveness of monomer form of antibacterial substances and greater activity is observed in higher concentrations (5 MIC).

The variance of the data due to the effects of three substances (*B*, *R* and *T*) in two oxidation forms and two concentration levels, the effects of two bacteria (*E. coli* and *H. pylori*) and the effect of one structure (monomer) is described by a three-way GEMANOVA term. What is not explained solely by this term is the effect of *β-myrcene* and *S* in their polymer shape, two oxidation forms and two concentrations on the *S. aureus*. The combination of instances of factors, which is not exclusively modeled using a three-way term, embraces only eight data out of the total 120 (5 × 3 × 2 × 2 × 2) (recall that each datum is the area under the kill kinetics curve recorded at a particular combination of instances of factors). The variances of these eight data are explained by two terms, a five-way and a three-way term. However the five-way term contributes to the variability to a greater extent (see the scale of the concentration plot in Fig. 3e). Therefore the following conclusions can be made:

(a) *β-myrcene* and *S* species kill *S. aureus* better when they are in the monomer form rather than in the polymer form.
(b) Oxidized forms of *β-myrcene* and *S* are more effective than their non-oxidized forms in killing *S. aureus*.
(c) *β-myrcene* and *S* show greater activity against *S. aureus* when their concentration is higher (5 MIC).

To be able to describe the complex interactions within this data set, the variance of data is split into two multi-way terms using GEMANOVA; one describing the data without eight data points of the effects of *β-myrcene* and *S* in their polymer shape on *S. aureus*, and the other describing the rest of the data points. This suggests that there should be a difference between the raw data of these eight points and the remaining data points. To further investigate this, the raw data, that is, the calculated area under the kill kinetics curve of 120 experiments of the first replicate is given in Fig. 4.

In this graph, the experiments in which *S. aureus* is incubated with *β-myrcene* or *S* in their polymer shape are shown by open diamonds and the rest of the experiments with filled diamonds. As can be seen only the eight data points (open diamonds) which were not modeled exclusively by the three-way term using GEMANOVA model are different from the rest of the data points. Therefore as discovered by the GEMANOVA model, two multi-way terms (perhaps two phenomena) explain the variance of this system.

To show the significance of the differences revealed between the instances of factors using the GEMANOVA model, it is possible to compare the estimated parameters using

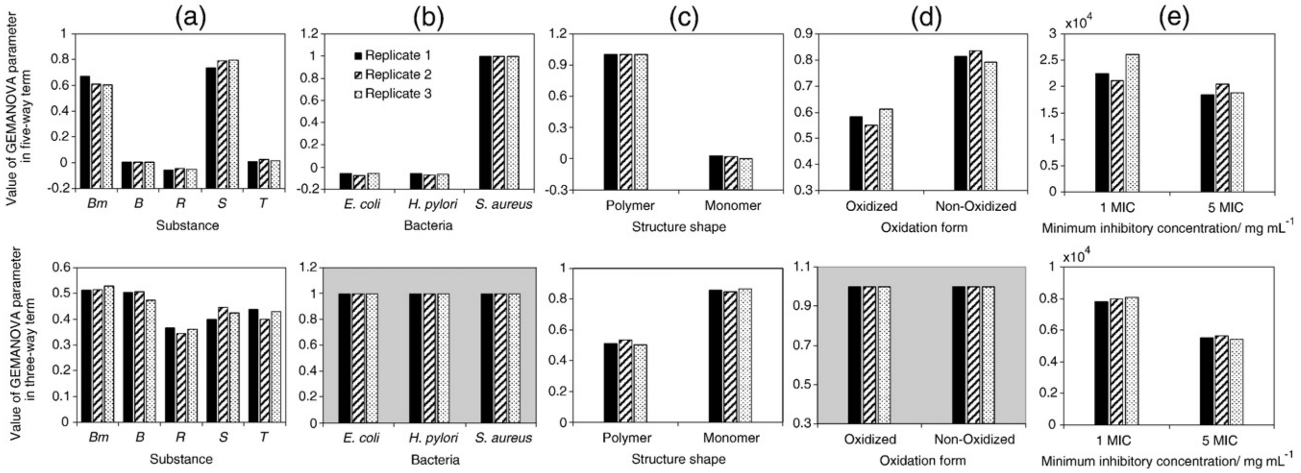

Fig. 5. Parameters of five- and three-way terms of GEMANOVA model of data from three independent replicates. Five profiles of (a) substance, (b) bacteria, (c) structure, (d) oxidation form and (e) concentration from the data sets of size (5 × 3 × 2 × 2 × 2) are shown. The plots of the GEMANOVA parameters for bacteria and oxidation form which are given unity in the three-way term are shown with a grey background.

GEMANOVA models of independent data sets [10]. In this work three independent replicates of the data were modeled individually using GEMANOVA. The results which are shown in Fig. 5 indicate a high agreement between the estimated parameters of different replicates. The similarity observed between the loading plots of independent data sets in this figure suggests that the GEMANOVA model is quite robust.

To obtain statistically sound conclusions regarding the significance of the differences between the instances of factors the Student's *t* test [17] was applied to the parameters estimated by GEMANOVA. Only the difference between *Bm* and *B* and the difference between *S* and *T* shown in Fig. 5(a) for three-way term were found to be non-significant at the 95% confidence level. The rest of the observed differences amongst the instances of factors were significant.

4.3. Comparison of ANOVA with GEMANOVA

While classic ANOVA describes the complex interactions within these data using several linear and two-to five-order interaction terms, GEMANOVA explains the entire variance using only two multi-way terms, which are easy to interpret. To compare the performance of the GEMANOVA model with that of ANOVA four figures of merits of RMSEC, RMSECP, r^2 and number of estimated parameters were used. The figures of merit were averaged over the three ANOVA models and then compared with the average of those in the three GEMANOVA models of three replicates. The result is shown in Table 1. These results were also compared with the average of the correlation between the raw data of replicates which was 0.97 in this experiment and the average of the residual error of repeating the experiment (RMSEP = 763).

As can be seen the correlation coefficients of the two methods of ANOVA and GEMANOVA are identical (0.93) and compatible with the actual correlation coefficient between replicates (0.97). Regarding the RMSEC and RMSEP values in Table 1 lower residual errors are returned using the classic ANOVA model when compared with that of GEMANOVA; however when the results are compared to the intrinsic error of this experiment, 763, we conclude that the ANOVA model has overfitted the data. The residual data shown in Table 1 have not been corrected for the number of parameters used in each model. In summary, the parsimonious GEMANOVA model has only 4% (=23/516 × 100) of the parameters of an equivalent ANOVA model, but has acceptable residual error and does not overfit the data. Therefore in this case GEMANOVA is clearly preferable to classic ANOVA.

5. Conclusion

Classic analysis of variance (ANOVA) was compared with GEMANOVA to interpret the kill kinetics data of the effect of new antibacterial agents with different concentrations, structures and oxidation forms on three species of bacteria. A large number of two-to five-order interaction terms (29 out of 31) were found to be significant using ANOVA, requiring up to 566 parameters in the resulting model. Due to the presence of complex interactions amongst the studied factors the interpretation of classic ANOVA models was impossible in this case. GEMANOVA on the other hand resulted in robust models which conformed to the actual structure of the data. The main sources of variance in the kill kinetics experiments of this study were shown to be due to the interaction between terms, rather than the main effects. Using GEMANOVA the interactions within this system were divided into two parts. In the first part the interaction between a subgroup of factors 'substance', 'bacterium' and 'structure' with all other factors was modeled using two multi-way terms of which the three-way term dominated. This term described the interaction between the substance subgroup of *B*, *R* and *T* with all other factors used in this term, interaction of bacteria subgroup of *E. coli* and *H. pylori* with all other factors used in this term and the interaction of the monomer shape of all substances with all other factors used in this term. In the second part the interaction between the substance subgroup of *β-myrcene* and *S* in their polymer form with the other two factors of oxidation form and concentration for the bacterium species of *S. aureus* was modeled using two multi-way terms of which the five-way term dominated. Division of the substance, bacterium and structure factors into two subgroups using GEMANOVA suggests that different bacteria display different resistance toward different substances in different structure shapes. In fact different behavior of *E. coli* and *H. pylori* which are both gram negative from *S. aureus* which is a gram positive bacterium can be explained by different structure of cell walls in these two groups of bacteria [18]. This study is the first application of GEMANOVA in the field of microbiology and the first application of GEMANOVA in which more than one multi-way term is required to give an interpretable model.

Acknowledgement

DE thanks the Australian Government for an International Postgraduate Research Scholarship.

References

[1] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke (Eds.), Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, 1997.

[2] A. Smilde, R. Bro, P. Geladi, Multi-Way Analysis with Applications in the Chemical Sciences, Wiley, Chichester, 2004.

[3] H.F. Gollob, Psychometrika 33 (1968) 73–115.

[4] V. Hegemann, D.E. Johnson, Technometrics 18 (1976) 273–281.

[5] J. Mandel, Technometrics 13 (1971) 1–18.

[6] J.R. Kettenring, A Case Study in Data Analysis, Proceedings of a Symposium in Applied Mathematics, 1983, p. 105.

[7] R. Bro, Multi-way Analysis in the Food Industry: Models, Algorithms and Applications, Department of Dairy and Food Science, Royal Veterinary and Agricultural University, Amsterdam, 1998.

[8] R. Bro, H. Heimdal, Chemometr. Intell. Lab. Syst. 34 (1996) 85–102.

[9] H. Heimdal, R. Bro, L.M. Larsen, L. Poll, J. Agric. Food Chem. 45 (1997) 2399–2406.

[10] R. Bro, M. Jakobsen, J. Chemometr. 16 (2002) 294–304.

[11] L.D. Nannerup, M. Jakobsen, F. van den Berg, J.S. Jensen, J.K.S. Moller, G. Bertelsen, Meat Sci. 68 (2004) 577–585.

[12] D.B. Hoellman, M.A. Visalli, M.R. Jacobs, P.C. Appelbaum, Antimicrob. Agents Chemother. 42 (1998) 857–861.

[13] H.G. Preuss, B. Echard, M. Enig, I. Brook, T.B. Elliott, Mol. Cell. Biochem. 272 (2005) 29–34.

[14] M.S. Sharifi, Fractionations and Analysis of Trunk Exudates from *Pistacia* genus in Relation to Antimicrobial Activity, College of Health and Science, University of Western Sydney, Sydney, 2006, p. 276.

[15] R.A. Newmark, R.N. Majumdar, J. Polym. Sci., A, Polym. Chem. 26 (1988) 71–77.

[16] K.J. Van den Berg, J. Van der Horst, J.J. Boon, O.O. Sudmeijer, Tetrahedron Lett. 39 (1998) 2645–2648.

[17] D.B. Hibbert, J.J. Gooding, Data Analysis for Chemistry An Introductory Guide for Students and Laboratory Scientists, Oxford University Press, New York, 2005.

[18] L.L. Silver, Biochem. Pharmacol. 71 (2006) 996–1005.