

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

**A Comparison of Univariate and Multivariate Statistical and Data Mining Approaches to the Behavioural and Biochemical Effects of Vestibular Loss Related to the Hippocampus**

**A thesis submitted in partial fulfilment of the requirements of the MAppStat in Applied Statistics, Massey University, Manawatu**

**Paul F. Smith**

**2013**

To my wife Cynthia and my cats, Max, Poppy and Chloe, all of whom have had to endure my journey into statistics over the last 8 years.

## **Acknowledgements**

First and foremost, I would like to thank my supervisor, Professor Stephen Haslett, for agreeing to supervise me and for his support and guidance throughout my thesis. His incisive analytical mind has been an inspiration to me. I would also like to thank Dr. Siva Ganesh, who stimulated my interest in data mining with his 4<sup>th</sup> year course on the subject.

This thesis would not have been possible without the data sets, and I thank Dr. Yiwen Zheng, Lucy Stiles, Irene Cheung and Georgina Wilson for meticulously collecting the data used in this thesis, and also to Irene for the use of her maze photos. I particularly want to thank Yiwen for her endless encouragement and for taking care of so many things while I was doing this.

I have to thank my wife, Cynthia, for carefully reading and commenting on the entire thesis draft, for asking questions to help clarify my writing and for her unwavering support in my pursuit of a statistical education.

Lastly, I want to thank the Head of our Dept. of Pharmacology and Toxicology at the University of Otago School of Medical Sciences, Associate Professor Rhonda Rosengren, for her continuous support and encouragement to pursue the MAppStat.

## Table of Contents

### Abstract

- 1.0 Introduction
- 1.1 The vestibular system
- 1.2 Vestibular contributions to cognition and emotion
- 1.3 Data sets used in this thesis
- 1.4 Literature review and rationale for statistical approach
  - 1.4.1 Univariate statistical methods
  - 1.4.2 Fixed versus random effects
  - 1.4.3 General Linear Model assumptions
    - 1.4.3.1 Normality assumption
    - 1.4.3.2 Homogeneity of variance assumption
    - 1.4.3.3 Compound symmetry or sphericity assumption
    - 1.4.3.4 Repeated measures ANOVAs with unbalanced repeated measures designs and missing data
    - 1.4.3.5 Alternatives to repeated measures ANOVAs: Linear mixed model analysis
  - 1.4.4 Multivariate statistical methods
    - 1.4.4.1 Multivariate ANOVA (MANOVA) and linear discriminant analysis (LDA)
      - 1.4.4.1.1 Multivariate normality
      - 1.4.4.1.2 Homogeneity of the covariance matrices
      - 1.4.4.1.3 Number of variables
      - 1.4.4.1.4 Example of LDA in neuroscience
    - 1.4.4.2 Support vector machines
    - 1.4.4.3 Cluster analyses
    - 1.4.4.4 Multiple linear regression
    - 1.4.4.5 Random forest regression and classification
- 2.0 Methods
  - 2.1 Behavioural study
    - 2.1.1 Animals
    - 2.1.2 Surgery

- 2.1.3 Drugs
- 2.1.4 Behavioural testing
  - 2.1.4.1 Open field maze (OFM)
  - 2.1.4.2 Elevated plus maze (EPM)
  - 2.1.4.3 Elevated T maze (ETM)
  - 2.1.4.4 Spatial forced alternation T maze (STM)
- 2.1.5 Statistical analyses
  - 2.1.5.1 Sample sizes
  - 2.1.5.2 Open field maze
  - 2.1.5.3 Elevated plus maze
  - 2.1.5.4 Elevated T maze
  - 2.1.5.5 Spatial forced alternation T maze (STM)
  - 2.1.5.6 Multivariate statistical and data mining analyses
- 2.2 Biochemical study
  - 2.2.1 Animals
  - 2.2.2 Surgery
  - 2.2.3 Sample sizes
  - 2.2.4 T maze training
  - 2.2.5 Tissue preparation
  - 2.2.6 Western blotting
  - 2.2.7 Statistical analyses
- 3.0 Results
  - 3.1 Behavioural study: Univariate statistical analyses
    - 3.1.1 Open field maze results
      - 3.1.1.1 Distance travelled and velocity in the OFM
      - 3.1.1.2 Zone activity in the OFM
        - 3.1.1.2.1 Effects of buspirone on zone activity
        - 3.1.1.2.2 Effects of FG-7142 on zone activity
      - 3.1.1.3 Pre-drug rearing in the OFM
        - 3.1.1.3.1 Effects of buspirone on rearing
        - 3.1.1.3.2 Effects of FG-7142 on rearing
    - 3.1.2 Elevated plus maze
    - 3.1.3 Elevated T maze

	3.1.3.1	Pre-drug performance
	3.1.3.2	Effects of buspirone on avoidance and escape
	3.1.3.3	Effects of FG-7142 on avoidance and escape
	3.1.4	Spatial T maze results
	3.1.4.1	Pre-drug performance
	3.1.4.2	Effects of buspirone and FG-7142 on spatial memory
3.2		Behavioural study: Multivariate statistical and data mining analyses
	3.2.1	MANOVA
	3.2.2	Linear discriminant analysis
	3.2.3	Random forests to predict group membership
	3.2.4	Support vector machines to predict group membership
	3.2.5	Cluster analyses
	3.2.6	Multiple linear regression
	3.2.7	Random forest regression
3.3		Biochemical study: Multivariate statistical analyses
	3.3.1	Protein expression in the hippocampus at 24 hs, 72 hs and 1 week post-BVD
	3.3.2	Protein expression in the hippocampus at 1 month post-BVD
	3.3.3	Protein expression in the hippocampus at 6 months post-BVD
4.0		Discussion
	4.1	Scientific interpretation
	4.1.1	Behavioural data
	4.1.2	Neurochemical data
	4.2	Statistical discussion
	4.2.1	Behavioural data
	4.2.3	Neurochemical data
	4.3	General conclusion
5.0		References
6.0		Appendices
	6.1	Appendix 1
	6.2	Appendix 2

6.3 Appendix 3

6.4 Appendix 4

Published papers associated with this thesis:

Zheng, Y., Cheung, I., Smith, P.F. (2012) Performance in anxiety and spatial memory tests following bilateral vestibular loss in the rat and effects of anxiolytic and anxiogenic drugs. *Behavioural Brain Research*. 235; 21-29.

Smith, P.F. (2012) Statistical analysis in pharmacology is not always BO. *Trends in Pharmacological Sciences*. 33; 565-566.

Smith, P.F. (2012) A note on the advantages of using linear mixed model analysis with maximal likelihood estimation over repeated measures ANOVAs in psychopharmacology: Comment on Clark et al. (2012). *Journal of Psychopharmacology*. 26; 1605-1607.

Zheng, Y., Wilson, G., Stiles, L., Smith, P.F. (2013) Glutamate receptor subunit and calmodulin kinase II expression in the rat hippocampus, with and without T maze experience, following bilateral vestibular deafferentation. *PLoS One*. 8(2); e54527. doi:10.1371/journal.pone.0054527, pp. 1-10.

Smith, P.F., Haslett, S.J., Zheng, Y. (2013) A multivariate statistical and data mining analysis of spatial memory-related behavior following bilateral vestibular deafferentation in the rat. *Behavioural Brain Research*. 246; 15-23.

## List of Figures

- Figure 1: The peripheral vestibular system in the inner ear.
- Figure 2: Schematic diagram illustrating the connections between vestibular sensation in the inner ear, the generation of eye movement (gaze stabilization, VOR), postural reflexes (posture and balance, VSR) and the estimation of self-motion.
- Figure 3: Balance control.
- Figure 4: Rose diagram showing the initial heading angles of sham and bilateral vestibular deafferentation (BVD) animals in darkness in a foraging task.
- Figure 5: Putative pathways from the brainstem vestibular nucleus to the hippocampus.
- Figure 6: Example of place cell dysfunction in the rat hippocampus following bilateral vestibular lesions.
- Figure 7: Normal Q-Q plots showing an example of deviation from normality for the CA2/3 data.
- Figure 8: A natural log transformation of the CA2/3 data partially resolves the violation of the normality assumption.
- Figure 9: Raw CA1 data showing non-normal distribution.
- Figure 10: Bootstrapping with 1000 samples of  $n = 8$  from the CA1 data.
- Figure 11: Gradual decrease in ‘spontaneous nystagmus’ following unilateral vestibular lesions in guinea pigs treated with adrenocorticotrophic factor 4-10 (ACTH-(4-10)) or its vehicle.
- Figure 12: Mean escape latency in sec in the 3 avoidance trials in the elevated T maze for the BVD and sham groups.
- Figure 13: Examples of covariance matrix structures.
- Figure 14: The arginine metabolic pathway.
- Figure 15: Example of a support vector machine (SVM) classification employing a radial basis function to separate iris varieties based on petal width and length.
- Figure 16: Dendrograms showing the relationship between the expression of the different neurochemical variables in the young and aged vestibular nucleus complex.
- Figure. 17: Diagnostic plots for L-citrulline following multiple linear regression (MLR) showing residuals versus fitted values, normal Q-Q, scale location and residuals versus leverage plots.
- Figure 18: Comparison of the adjusted  $R^2$  values (MLRs) and variance explained values (random forest regressions; RFRs), and the residual mean square error (RSE)

values, for the MLRs and RFRs for the neurochemical variables from Liu et al. (2010).

Figure 19: The open field maze (OFM) with its 3 zones.

Figure 20: The elevated plus maze (EPM).

Figure 21: The elevated T maze (ETM).

Figure 22: The spatial T maze (STM)

Figure 23: A Spearman correlation matrix for a subset of the behavioural variables.

Figure 24: Spearman correlation clusters showing the relationship between each behavioural variable and every other behavioural variable.

Figure 25: Mean distance travelled (A) and velocity (B) in the OFM.

Figure 26: Comparison of the raw zone data with the bootstrapped zone data for the sampling distribution of the mean.

Figure 27: Mean % time spent in the inner/middle and outer zones of the OFM for BVD and sham animals over a 1, 5 and 10 min period pre-drug (A) and in response to buspirone (B) or FG-7142 (C).

Figure 28: Mean frequency of supported rearing in the OFM in BVD and sham groups pre-drug and in response to buspirone or FG-7142.

Figure 29: Mean frequency of unsupported rearing in the OFM in BVD and sham groups pre-drug and in response to buspirone or FG-7142.

Figure 30: Mean duration of supported (A) and unsupported (B) rearing in the OFM in BVD and sham groups pre-drug and in response to buspirone or FG-7142.

Figure 31: Mean % time in the open arms (A) and distance travelled (B) for the BVD and sham groups pre-drug and in response to buspirone or FG-7142.

Figure 32: Mean avoidance (A) and escape latency (B) in the 3 trials in the ETM for the BVD and sham groups pre-drug and in response to buspirone or FG-7142.

Figure 33: Mean % correct responses in the STM for the BVD and sham animals before any drug treatment (A) and in response to buspirone or FG-7142 (B).

Figure 34: The variables in order of importance for the random forest (RF) model.

Figure 35: The out-of-bag (OOB), BVD and sham group error rates for the RF model as a function of the number of trees.

Figure 36: Cluster analysis of the behavioural data.

Figure 37: Normality plot for the residuals of the MLR analysis.

Figure 38: Residual analysis for the best subsets regression model.

Figure 39: Variables in order of importance for the RFR.

Figure 40: Error for the RFR as a function of the number of trees.

Figure 41: Predicted versus observed values for ln % correct for the test data in the RFR.

Figure 42: Percentage of correct choices in the T maze task for bilateral vestibular deafferentation (BVD) and sham surgery controls at 3 weeks (A), 3 months (B) and 5 months (C) post-op.

Figure 43: Mean protein expression for NR1, NR2A, NR2B, GluR1, GluR2, GluR3 and GluR4 receptors and CamKII $\alpha$  and pCamKII $\alpha$  in the CA1, CA2/3 and DG regions of the hippocampus at 24 hs, 72 hs or 1 week following BVD and sham surgery.

Figure 44: Mean normalized density of expression of NR1, NR2B, GluR1, GluR2, GluR3, CaMKII $\alpha$  and pCaMKII $\alpha$  in the CA1, CA2/3 and DG regions of the hippocampus at 6 months following BVD or sham surgery for animals trained in a T maze or not trained in a T maze.

Figure 45: Example of western blots for CaMKII $\alpha$  and pCaMKII $\alpha$  in CA2/3 for the BVD and sham animals that received T maze training or no T maze training at 6 months post-op.

Figure 46: Spearman correlation analysis for the different glutamate receptor subunits, CaMKII $\alpha$  and pCaMKII $\alpha$ , in the 3 hippocampal subregions at 6 months post-op., with the BVD and sham animals and T maze-trained and non-T maze-trained animals' data pooled.

Figure 47: Spearman correlation cluster analysis for the different glutamate receptor subunits, CaMKII $\alpha$  and pCaMKII $\alpha$  in the 3 hippocampal subregions at 6 months post-op., with the BVD and sham animals and T maze-trained and non-T maze-trained animals' data pooled.

Figure 48: Variables in order of importance for the prediction of surgical group using a RF classification analysis.

Figure 49: OOB, BVD and sham classification error for the prediction of surgical group using a RF classification analysis.

Figure 50: Variables in order of importance for the RF classification for T maze training and no T maze training.

Figure 51: OOB, T maze training and no T maze training error rates as a function of the number of trees.

Figure 52: Cluster analysis of all of the neurochemical data in CA1 (A), CA2/3 (B) and the DG (C), showing individual animals according to surgery.

Figure 53: Cluster analysis of the neurochemical data at 6 months post-op. in CA1 (A), CA2/3 (B) and the DG (C), showing individual animals according to training (T maze or no T maze).

Figure 54-59: Other methods of cluster analysis, such as between group (average) linkage, within group (average) linkage, nearest neighbour (single linkage), furthest neighbour (complete linkage) centroid clustering and median clustering, for the 8 variables from the behavioural data.

## **List of Tables**

- Table 1: Statistical analyses used by the relevant behavioural and neurochemical animal studies.
- Table 2: Example of Kolmogorov-Smirnov and Shapiro-Wilk tests of normality for hippocampal data from this thesis.
- Table 3: Normality tests of the data following ln transformation.
- Table 4: Levene's test of homogeneity of variance for the pre-drug locomotor velocity data set.
- Table 5: Levene's test of homogeneity of variance for the complete locomotor velocity data set.
- Table 6: Example of Mauchly's test of sphericity for the escape latency data, showing a significant violation of the sphericity assumption.
- Table 7: Example of Greenhouse-Geisser and other corrections for the violation of the assumption of sphericity in an ANOVA for the escape latency data.
- Table 8: Linear discriminant function coefficients from the Liu et al. (2010) study.
- Table 9: The order of the drugs administered to each animal during 4 repetitions of each behavioural task.
- Table 10: The order of the behavioural tests.
- Table 11: The results of the Kaiser-Meyer-Olkin Measure of Sampling Adequacy for principal component analysis (PCA).
- Table 12: Primary and secondary antibodies and their dilutions used in the western blotting experiments.
- Table 13: Kolmogorov-Smirnov and Shapiro-Wilk normality tests for the pre-drug locomotor distance data.
- Table 14: F and Levene's tests of homogeneity of variance for the pre-drug locomotor distance data.
- Table 15: Box's M test for the equality of the covariance matrices.
- Table 16: The standardized canonical discriminant function coefficients for the linear discriminant analysis (LDA) with all variables entered.
- Table 17: The classification matrix for the LDA with all variables entered.
- Table 18: The standardized canonical discriminant function coefficients for the stepwise LDA.
- Table 19: The classification matrix for the stepwise LDA.

Table 20: The confusion matrix for the RF analysis.

Table 21: Overall error for the RF model.

Table 22: Results of the support vector machine (SVM) analysis using the Gaussian Radial Basis kernel function.

Table 23: Behavioural variables used in the MLRs.

Table 24: Backward regression results for the MLR.

Table 25: ANOVA table for the backward regression MLR.

Table 26: Best subsets regression results.

Table 27: OOB and classification error rates for the RF analysis of the pooled 6 month neurochemical data for the prediction of training in a T maze.

Table 28: Example of the results of the normality tests for the 5 data sets for the locomotor distance data analysed in Section 3.1.1.1.

Table 29: Example of the results of the homogeneity of variance tests for the 5 data sets for the locomotor distance data analysed in Section 3.1.1.1.

Table 30: Example of the results of a 2 sample independent t test for the pre-drug locomotor distance data analysed in Section 3.1.1.1.

Table 31: Example of Mauchly's test of sphericity for the locomotor distance data analysed in Section 3.1.1.1.

Table 32: Example of a repeated measures ANOVA, in this case, for the locomotor distance data analysed in Section 3.1.1.1.

Table 33: Example of a repeated measures ANOVA, in this case, for the locomotor distance data analysed in Section 3.1.1.1.

Table 34: Example of a linear mixed model (LMM) analysis, using an unstructured covariance matrix, for the zone data analysed in Section 3.1.1.2.

Table 35: Example of a Bonferroni post-hoc test on avoidance latency, as reported in Section 3.1.3.2.

Table 36: Example of the MANOVA performed on the behavioural data, as reported in Section 3.2.1.

Table 37: Example of the homogeneity of variance tests performed on the behavioural data, as reported in Section 3.2.1.

Tables 38-42: Example of the stepwise LDA reported in Section 3.2.2.

Tables 43 and 44: Example of the MANOVA performed on the neurochemical data for CA1 at 6 months post-op, as reported in Section 3.3.3.

## List of Abbreviations

A3:	3 <sup>rd</sup> avoidance latency in the ETM
AIC:	Akaike's Information Criterion
AMPA:	$\alpha$ -amino-3-hydroxy-5-methyl-4-isoxazolepropionate
ANCOVA:	analysis of covariance
ANOVA:	analysis of variance
AR(1):	autoregressive order 1
ARMA:	autoregressive moving average
BIC:	Bayesian Information Criterion
Bus:	bupirone
BVD:	bilateral vestibular deafferentation
CA:	cluster analysis
CaMKII $\alpha$ :	calmodulin kinase II $\alpha$
CE:	cerebellum
CI:	confidence interval
DG:	dentate gyrus
Dist:	distance travelled in the OFM
E3:	3 <sup>rd</sup> escape latency in the ETM
EPM:	elevated plus maze
Epmdur:	duration of open arm entries in the EPM
Epmdist:	distance travelled in the EPM
Epmfreq:	frequency of open arm entries in the EPM
ETM:	elevated T maze
FG-7142:	N-methyl- $\beta$ -carboline-3-carboxamide
Ln IO:	ln of ratio of time spent in the inner to the outer zone of the OFM
Ln percent:	ln of percent correct in the STM
i.p:	intraperitoneal
LDA:	linear discriminant analysis
LDF:	linear discriminant function
LMM:	linear mixed model
LOO:	leave one out
MANOVA:	multivariate analysis of variance
MCAR:	missing completely at random
MAR:	missing at random
MLE:	maximum likelihood estimation

MLR:	multiple linear regression
MSE:	mean square error
MVA:	missing values analysis
NMDA:	N-methyl-D-aspartate
OFM:	open field maze
OOB:	out of bag
PCA:	principal component analysis
pCaMKII $\alpha$ :	phosphorylated calmodulin kinase II $\alpha$
QDA:	quadratic discriminant analysis
REML:	restricted maximal likelihood estimation
RF:	random forest
RFR:	random forest regression
ROC:	receptor operating characteristic
RSE:	residual mean square error
s.c:	subcutaneous
Sdur:	duration of supported rearing
Sfreq:	frequency of supported rearing
SN:	spontaneous nystagmus
SSE:	sum of squares for the error
SST:	sum of squares for the treatments
STM:	spatial T maze
SVM:	support vector machine
Udur:	duration of unsupported rearing
Ufreq:	frequency of unsupported rearing
UVD:	unilateral vestibular deafferentation
VIF:	variance inflation factor
VNC:	vestibular nucleus complex
VOR:	vestibulo-ocular reflex
VSR:	vestibulo-spinal reflex

## **Abstract**

Vestibular dysfunction is associated with a complex syndrome of cognitive and anxiety disorders. However, most studies have used simple univariate analyses of the effects of vestibular loss on behaviour and brain function. In this thesis, univariate statistical, and multivariate statistical and data mining approaches, to the behavioural and neurochemical effects of bilateral vestibular deafferentation (BVD), were compared. Using linear mixed model analyses, including repeated measures analyses of variance and analyses with the covariance structure of the repeated measures specified, rats with BVD were found to exhibit increased locomotor activity, reduced rearing and reduced thigmotaxis. By contrast, there were no significant differences between BVD and sham control animals in the elevated plus maze and the BVD animals exhibited a longer escape latency in the elevated T maze, with no change in avoidance latency. In the spatial T maze, the BVD animals demonstrated a significant decrease in accuracy compared to the sham control animals. Using linear discriminant analysis, cluster analysis, random forest classification and support vector machines, BVD animals could be distinguished from sham controls by their behavioural syndrome. Using multiple linear regression and random forest regression, the best predictors of performance in the spatial T maze were whether the animals had received a BVD or sham lesion, and the duration of rearing. In the neurochemical data set, the expression of 5-7 glutamate receptor subunits was measured in 3 different subregions of the rat hippocampus, at various times following BVD, using western blotting. In the 6 month group, half of the animals underwent training in a T-maze. Using multivariate analyses of variance, there was no significant effect of surgery for any hippocampal subregion. Linear discriminant analysis could not determine a linear discriminant function that could separate BVD from sham control animals. A random forest classification analysis was also unsuccessful in this respect. However, for the 6 month data set, T maze training had a significant effect independently of surgery. The results of these experiments suggest that BVD results in profound spatial memory deficits that are not associated with large changes in the expression of glutamate receptors in the hippocampus. The results of the multivariate statistical and data mining analyses, applied to both the

behavioural and neurochemical data sets, suggested that research in this field of neuroscience would benefit from analysing multiple variables in relation to one another, rather than simply conducting univariate analyses. Since the different behavioural and neurochemical variables do interact with one another, it is important to determine the nature of these interactions in the analyses conducted. However, this will require researchers to design experiments in which multiple variables can be measured under the one set of conditions.