www.csiro.au

Remote Analysis & Differential Privacy

Remote Analysis vs SDC for Business Data

Christine M O'Keefe
CSIRO Mathematics, Informatics and Statistics

1 July 2011

CSIRO

Overview

- Remote Analysis & Differential Privacy
  - Logistic regression
  - Other models
    - Work in progress – discussions with
      - James Chipperfield, Sebastien Lucie (Aust Bureau Statistics)
      - Steve Fienberg, Alessandro Rinaldo (CMU)

Headline conclusion:

can be better to add noise to something other than output

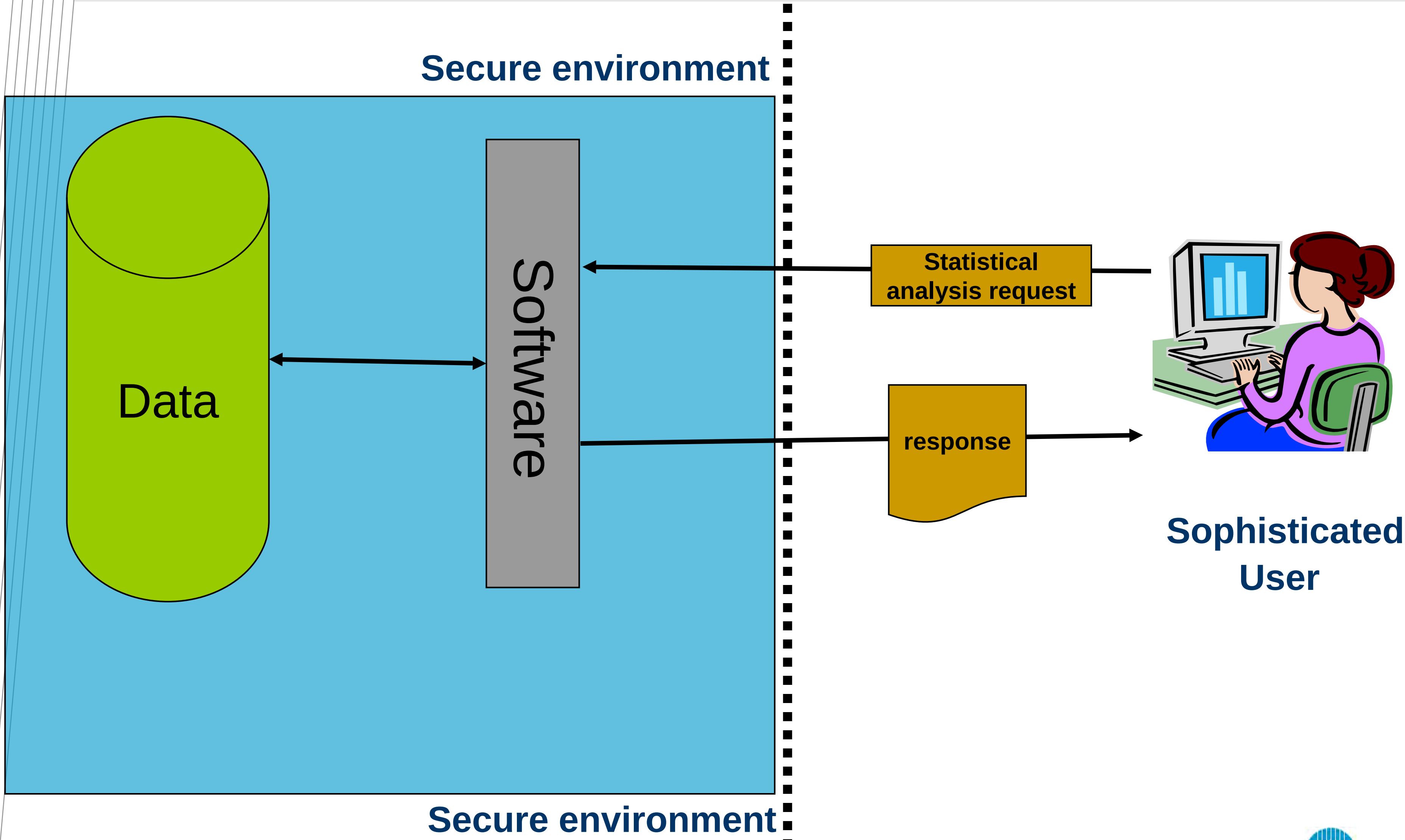
- Remote Analysis vs Statistical Disclosure Control
  - Business Data
    - Joint work with Natalie Shlomo (submitted)

Headline conclusion:

remote analysis (output perturbation) seems preferable…

Remote Analysis

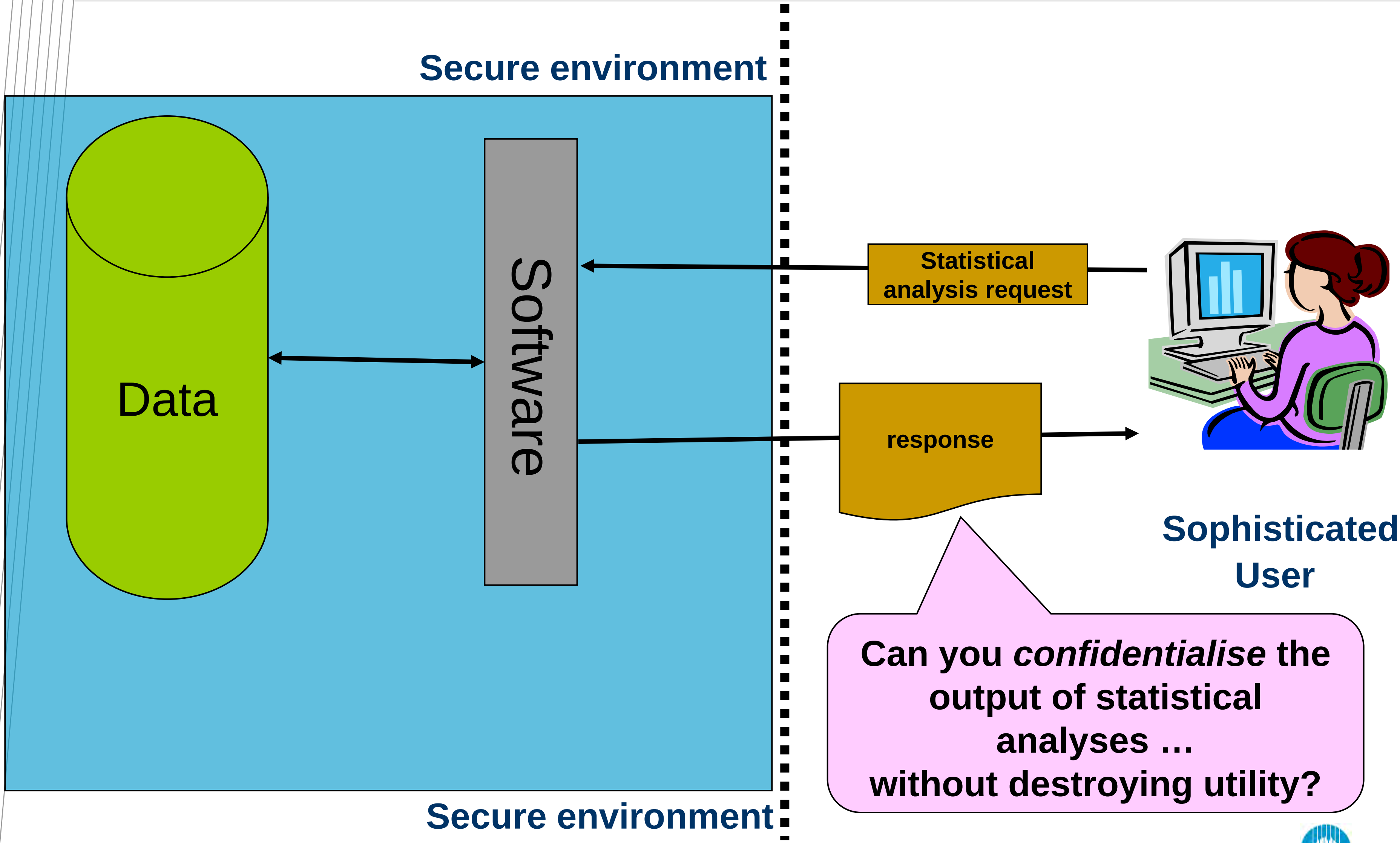
Remote Analysis – the basic scenario



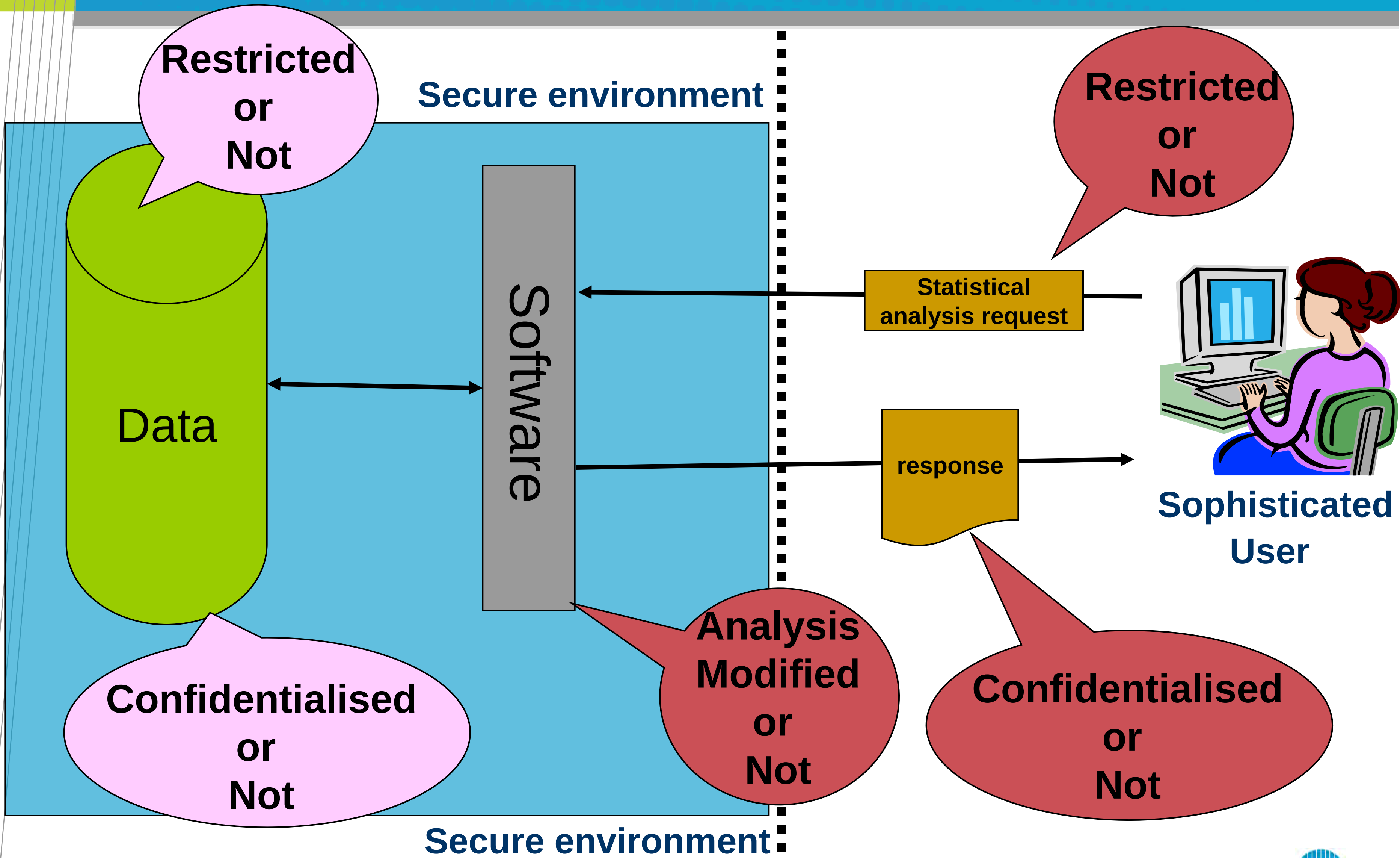
Scenarios where remote analysis may be useful

- Phase 1 investigations with low-risk ethical review – prior to applying for full access with full ethical review
  - Prepare before visiting data laboratory
  - Preliminary results for grant applications
  - Evidence that application for full access is worthwhile
- Restricted functionality may be sufficient in some situations
- Some data may be unavailable by other means
  - Business data
- Analyst activity can be easily logged for monitoring and audit

Remote Analysis – the $1M question

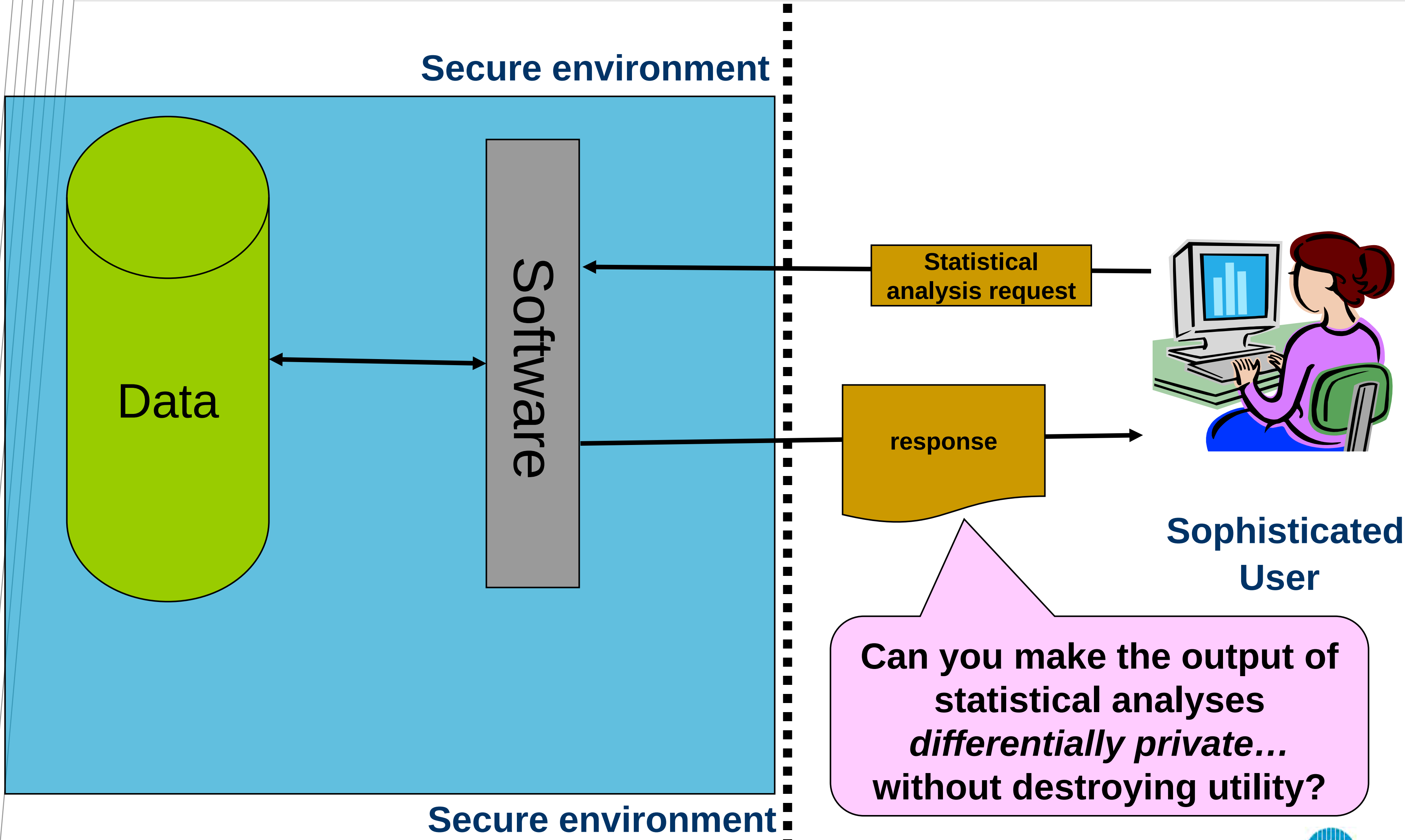


Remote Analysis – options for intervention

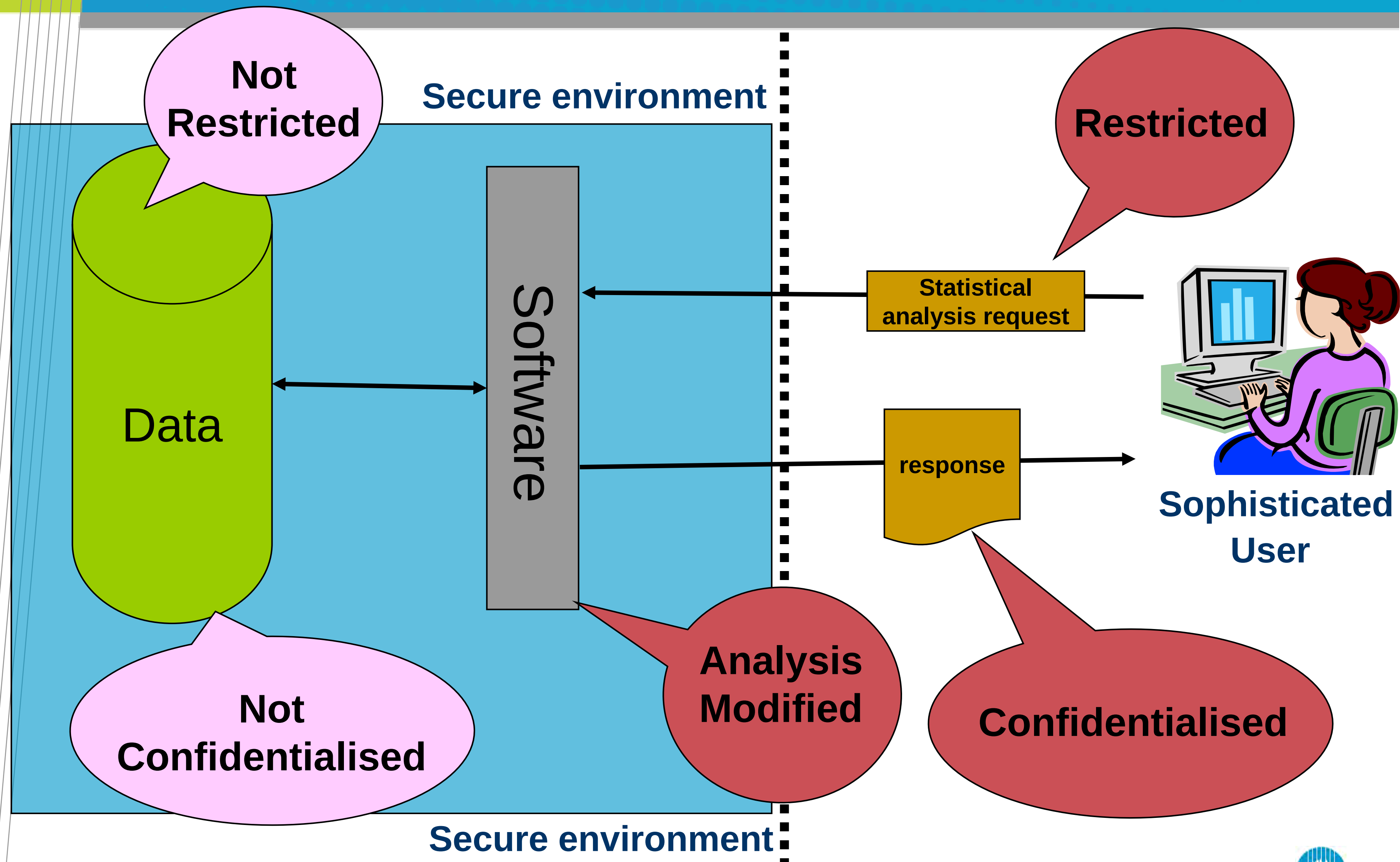


Differential Privacy

Differential Privacy – the $1M question



Differential Privacy– options for intervention



Example

Logistic Regression

Example – logistic regression

- Chaudhuri & Monteleoni 2008 - perturbing the estimate
  - Explanatory variable data space bounded by unit ball
  - Binary response variable

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \beta^T x_i)) + \frac{1}{2} \lambda \beta^T \beta$$

λ a regularising parameter

  - Function with output $\hat{\beta}$ has sensitivity 2/nλ
  - $\hat{\beta}$ + α where α ~ Lap(2/nλε) is ε-differentially private

- Smith 2008
  - Sample-and-aggregate the bias-corrected maximum likelihood estimate
  - Add sensitivity-calibrated noise

CSIRO

Example – logistic regression

- Chaudhuri & Monteleoni 2008 - perturbing the objective function
  - Explanatory variable data space bounded by unit ball
  - Binary response variable

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \log(1+\exp(-y_i \beta^T x_i)) + \frac{1}{2}\lambda\beta^T\beta + \frac{1}{n}\alpha^T\beta$$

where $\alpha \sim \mathrm{Lap}(2/\varepsilon)$, is ε-differentially private

- Perturbing something other than the “output”

Remote Analysis vs Statistical Disclosure Control for Business Data

Christine M O'Keefe and Natalie Shlomo
CSIRO Mathematics, Informatics and Statistics
Southampton Statistical Sciences Research Institute

1 July 2011

Outline

- Business data
  - Particular challenges
  - Current approaches
- Example: Sugar Farms Data
  - Statistical disclosure control
  - Remote analysis
  - Comparison
    - Exploratory data analysis
    - Linear regression
- Headline conclusion:

  remote analysis seems preferable

Business data

Business data - challenges

- Characteristic pattern of inclusion probabilities
  - Large enterprises always sampled - census
  - Medium-sized enterprises often sampled
  - Small enterprises seldom sampled
- Few variables
- Most variables continuous not discrete
- Most variable distributions highly skewed
- Common to have enterprises which are outliers on almost all variables

Example

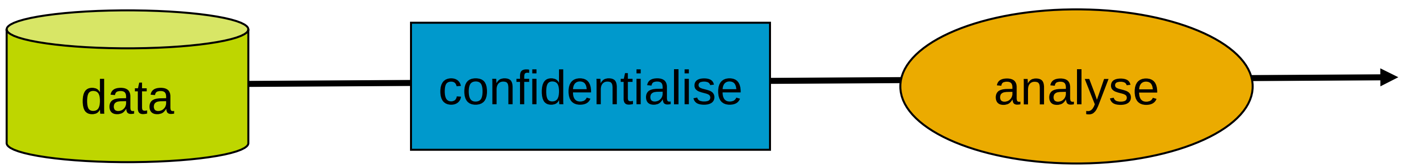
Sugar Farms Data

Sugar Farms Data

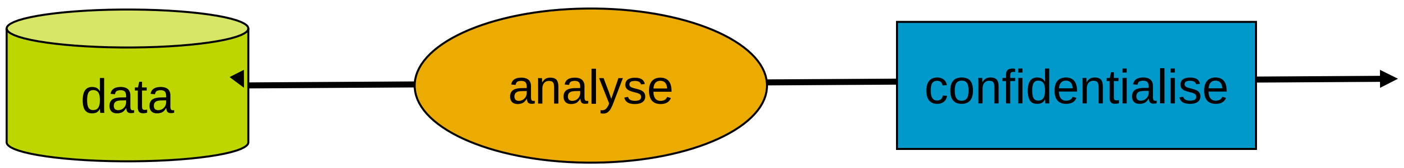
- 1982 survey of sugar cane industry in Queensland
  - Sample of 338 Queensland sugar cane farms
  - Stratified by region and size of quota – random within strata
- Variables - categorical
  - region = cane growing region (1, 2, 3 or 4)
- Variables – continuous
  - area = area under sugar cane
  - harvest = quantity of sugar cane harvest
  - receipts = receipts from sale of sugar cane
  - costs = costs of growing sugar cane
  - profit = receipts - costs
- Characteristic of business data
  - 5 farms receipts over $300K outliers on all continuous variables

Statistical Disclosure Control vs Remote Analysis

- Statistical Disclosure Control – input perturbation



- Remote Analysis – output perturbation



Sugar Farms data - SDC

- Delete five largest farms – outliers
- Region
  - Not disclosive
- Area
  - Key identifying – categorised into 6 groups
- Harvest, receipts, costs, profit
  - Random noise – preserving mean and covariance structure

Sugar Farms data – Remote Analysis

- Ensure each combination of variable values has sufficient data cases represented
  - Data aggregation
- Rounding and smoothing of results
- Risks associated with outliers
  - Minimised by use of robust methods
  - Data winsoring
- Sought to ensure that SDC and RA approaches have comparable disclosure risk
  - Identity disclosure through small cells
  - Attribute disclosure through distance from true values

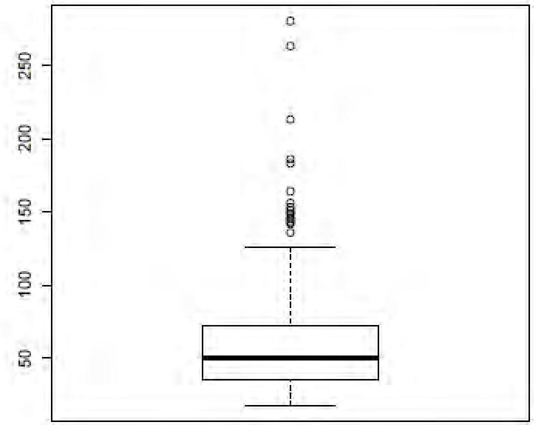
Example

Exploratory Data Analysis

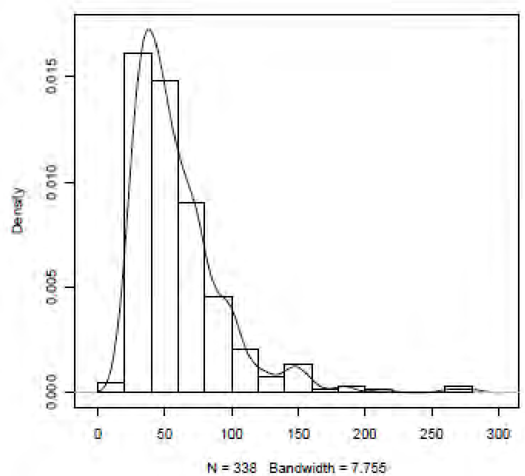
Univariate – area – unconfidentialised

Sugar Cane Area	
Minimum	18
1st Quartile	36
Median	51
Mean	60.25
3rd Quartile	73
Maximum	280
Standard Deviation	35.61062

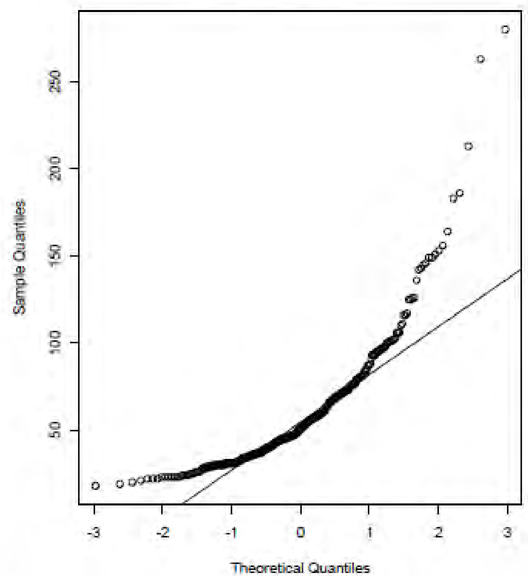
(a) Summary Statistics



(b) Box Plot



(c) Histogram and Density

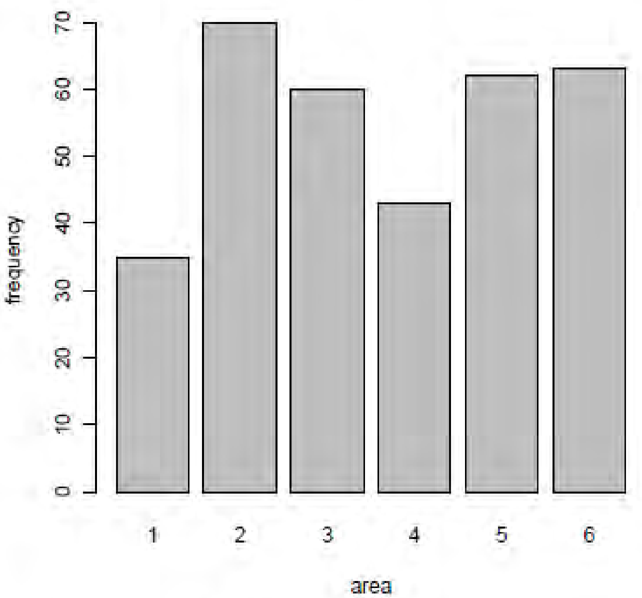


(d) Normal QQ-plot

Univariate – area – SDC

area		
	Category	**Frequency**
1	Up to 29	35
2	30 – 39	70
3	40 – 49	60
4	50 – 59	43
5	60 – 79	62
6	80 and over	63
	TOTAL	**333**

(a) Frequency Table

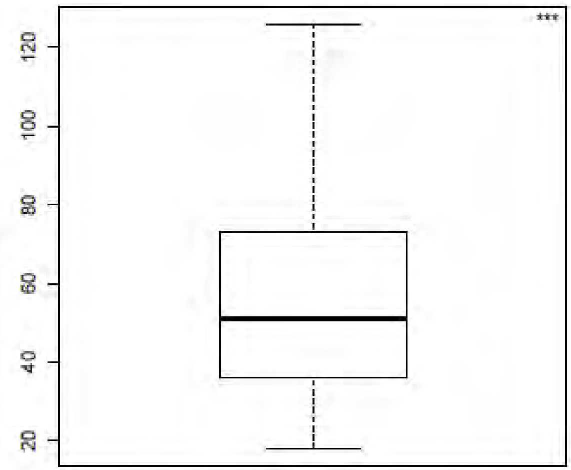


(b) Bar Chart

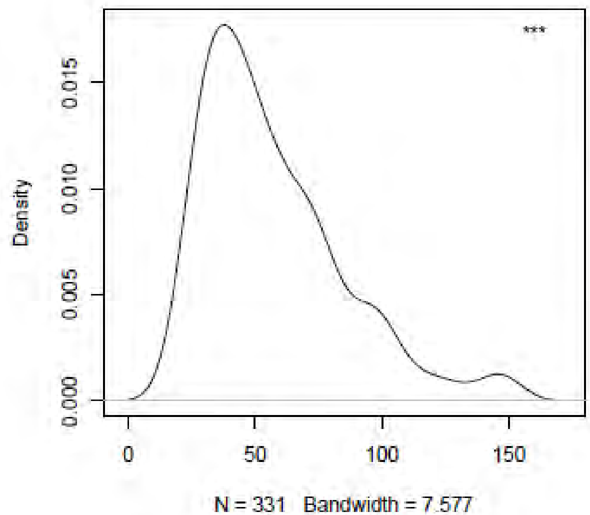
Univariate – area – remote analysis

Sugar Cane Area	
1st Quartile	35
Median	50
Mean	60.25
3rd Quartile	70
Standard Deviation	36

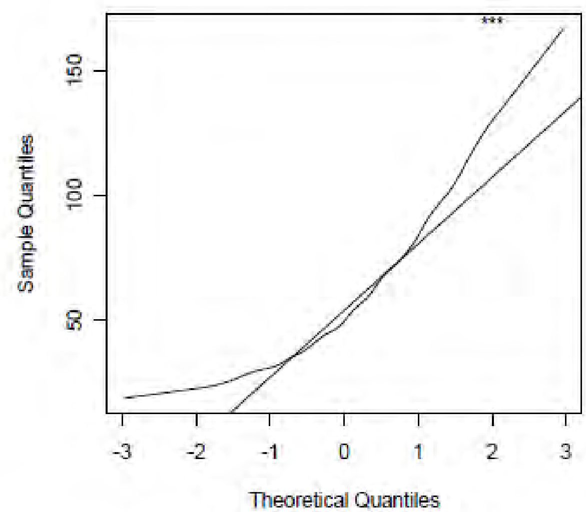
(a) Confidentialised Summary Statistics



(b) Confidentialised Box Plot



(c) Confidentialised Density Estimate

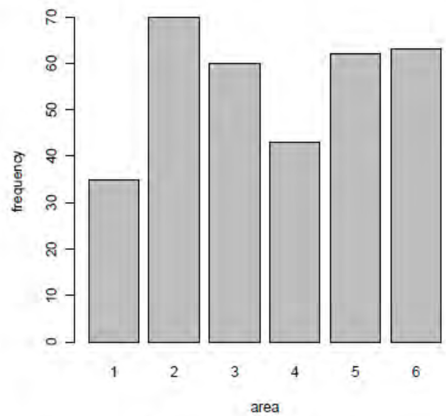


(d) Confidentialised QQ-Plot

Univariate – area – side by side

area		
	Category	**Frequency**
1	Up to 29	35
2	30 – 39	70
3	40 – 49	60
4	50 – 59	43
5	60 – 79	62
6	80 and over	63
	TOTAL	**333**

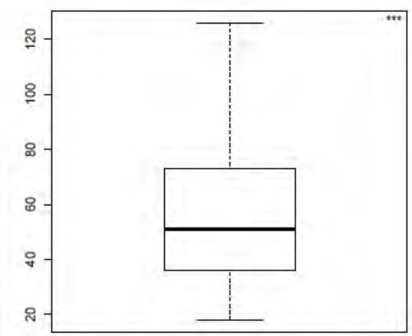
(a) Frequency Table



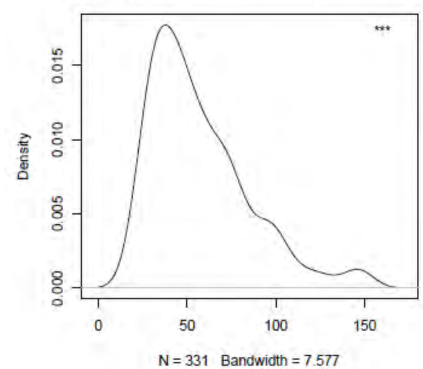
(b) Bar Chart

Sugar Cane Area	
1st Quartile	35
Median	50
Mean	60.25
3rd Quartile	70
Standard Deviation	36

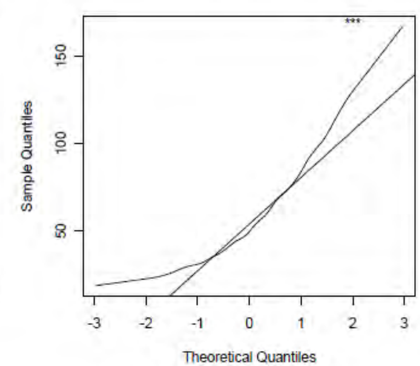
(a) Confidentialised Summary Statistics



(b) Confidentialised Box Plot



(c) Confidentialised Density Estimate

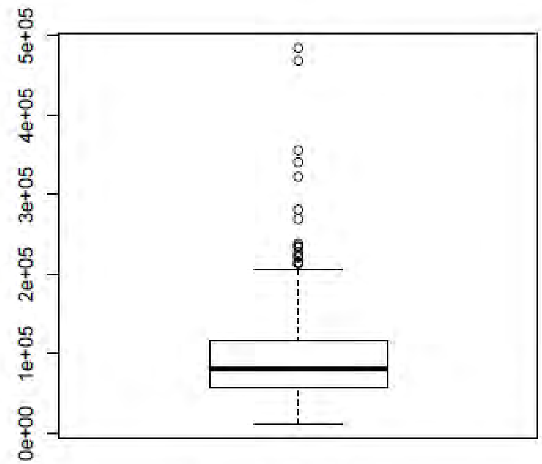


(d) Confidentialised QQ-Plot

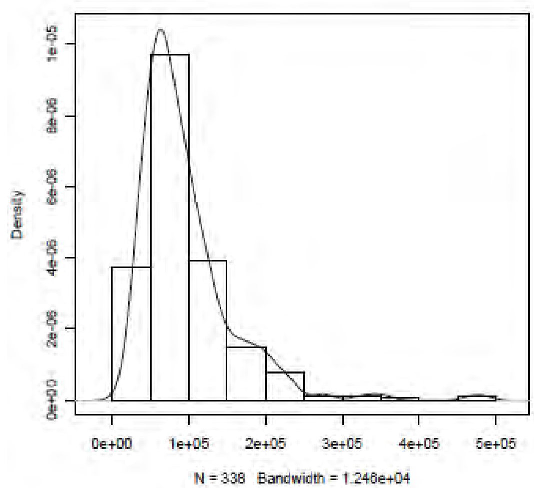
Univariate – receipts - unconfidentialised

	Receipts
Minimum	11703
1st Quartile	57607
Median	80391
Mean	95965
3rd Quartile	117062
Maximum	484346
Standard Deviation	61609.105256

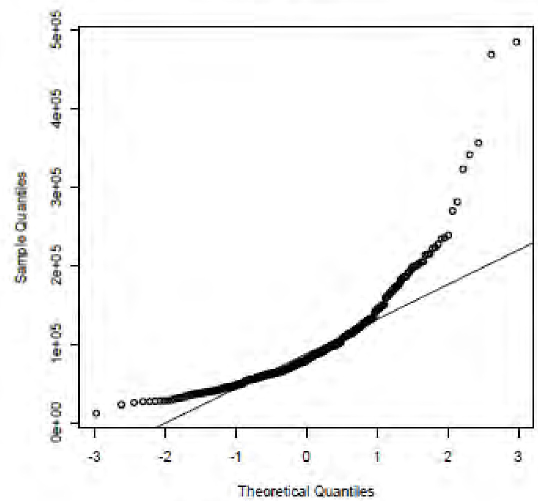
(a) Summary Statistics



(b) Box Plot



(c) Histogram and Density

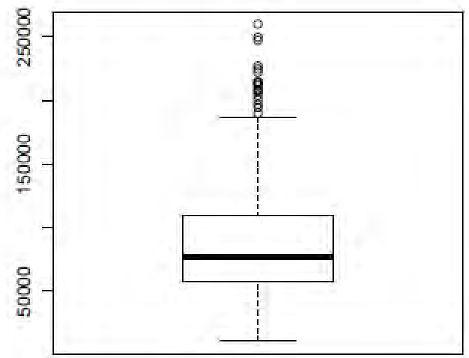


(d) Normal QQ-plot

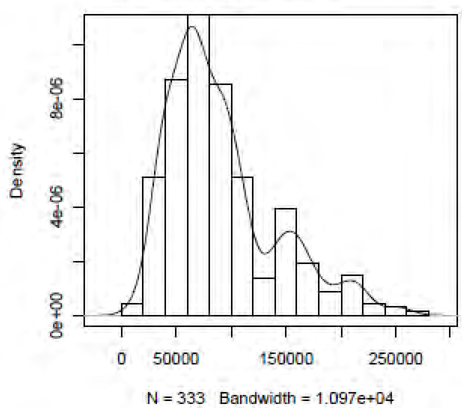
Univariate – receipts – SDC

	Receipts
No. observations	333
Minimum	11140
1st Quartile	57473
Median	77144
Mean	90643
3rd Quartile	109637
Maximum	260098
Standard Deviation	49214.06

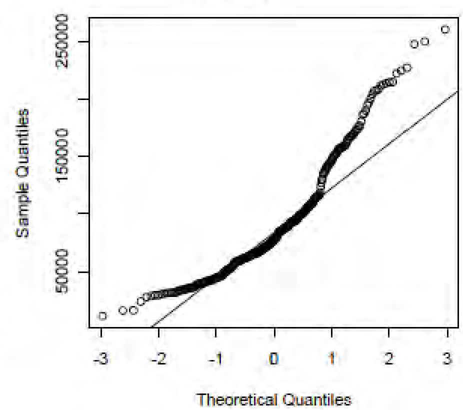
(a) Summary Statistics



(b) Box Plot



(c) Histogram and Density

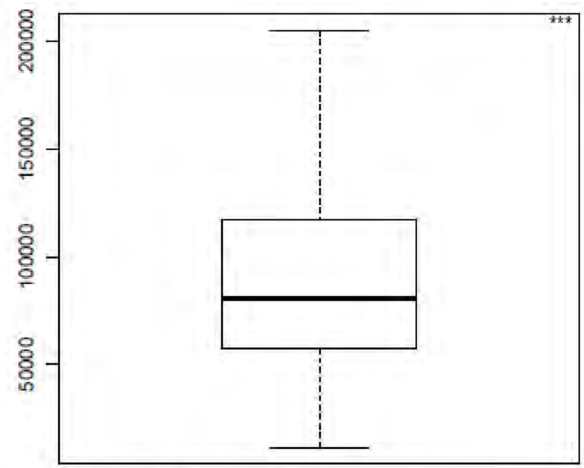


(d) Normal QQ-plot

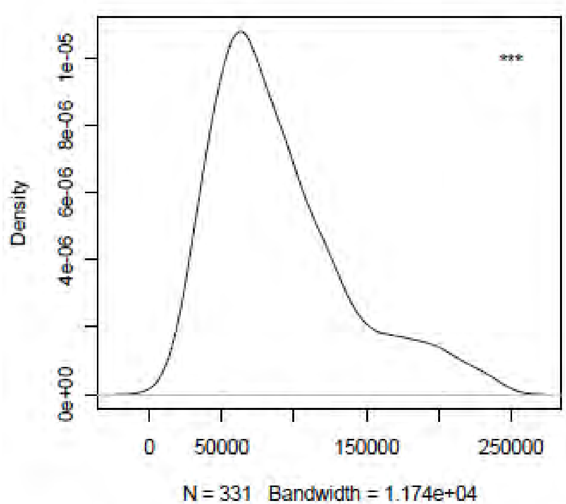
Univariate – receipts – remote analysis

	Receipts
1st Quartile	57600
Median	80400
Mean	96000
3rd Quartile	117100
Standard Deviation	61600

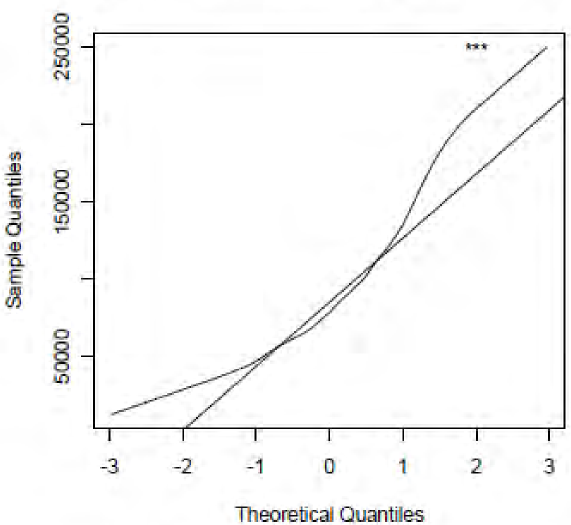
(a) Confidentialised Summary Statistics



(b) Confidentialised Box Plot



(c) Confidentialised Density Estimate

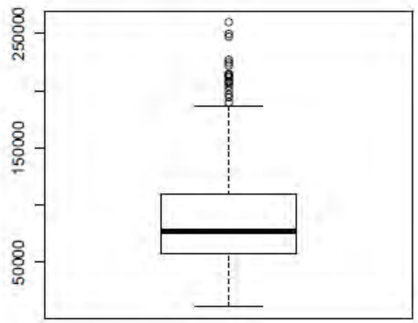


(d) Confidentialised QQ-Plot

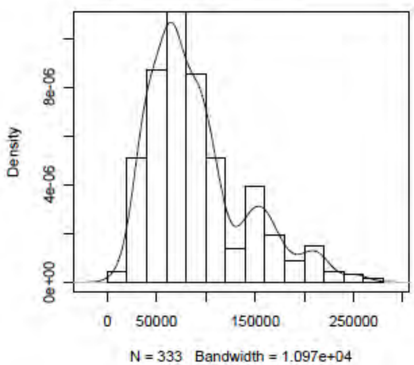
Univariate – receipts – side by side

	Receipts
No. observations	333
Minimum	11140
1st Quartile	57473
Median	77144
Mean	90643
3rd Quartile	109637
Maximum	260098
Standard Deviation	49214.06

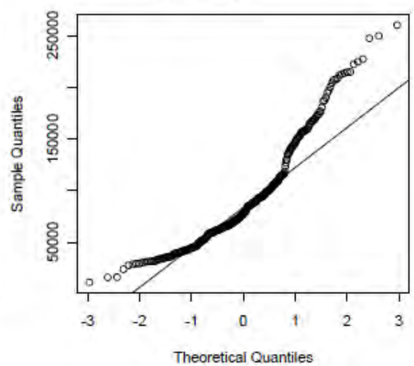
(a) Summary Statistics



(b) Box Plot



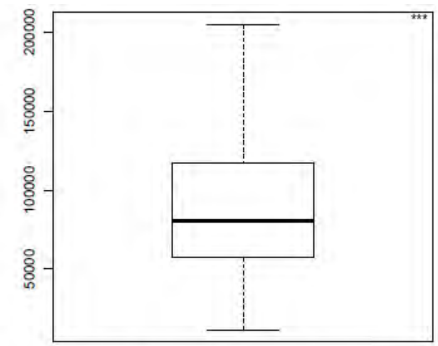
(c) Histogram and Density



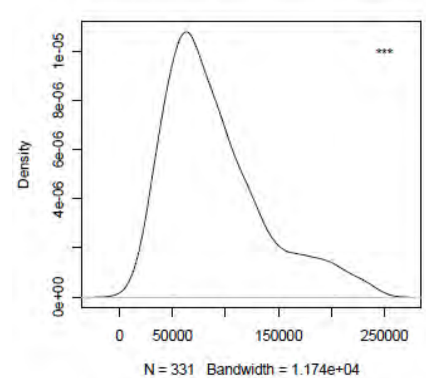
(d) Normal QQ-plot

	Receipts
1st Quartile	57600
Median	80400
Mean	96000
3rd Quartile	117100
Standard Deviation	61600

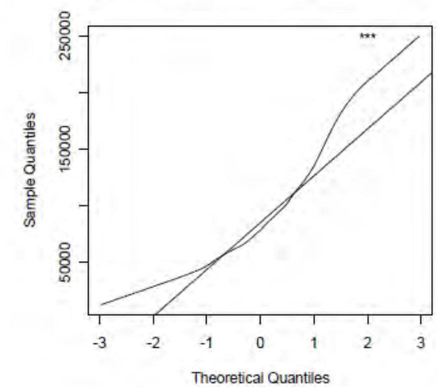
(a) Confidentialised Summary Statistics



(b) Confidentialised Box Plot

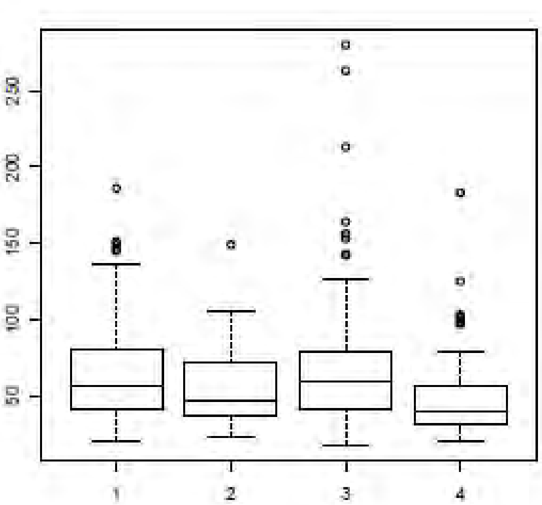


(c) Confidentialised Density Estimate

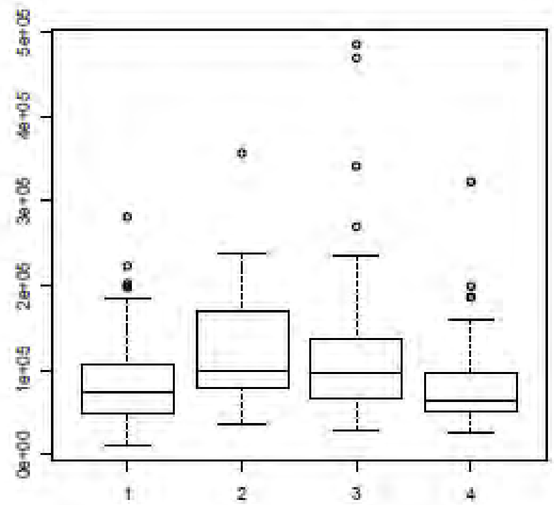


(d) Confidentialised QQ-Plot

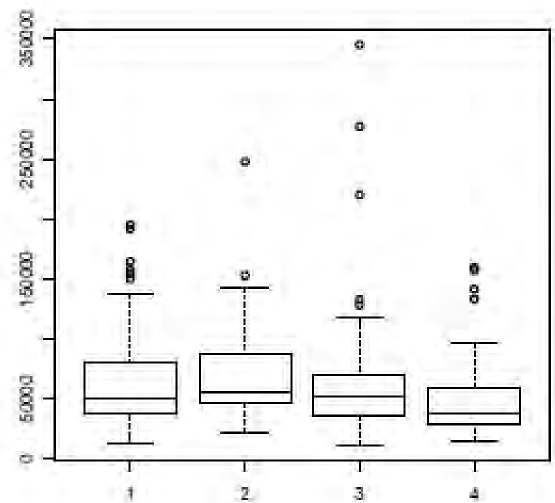
Bivariate - area, receipts, costs by region - unconfidentialised



(a) Box plots for area by region



(b) Box plots for receipts by region

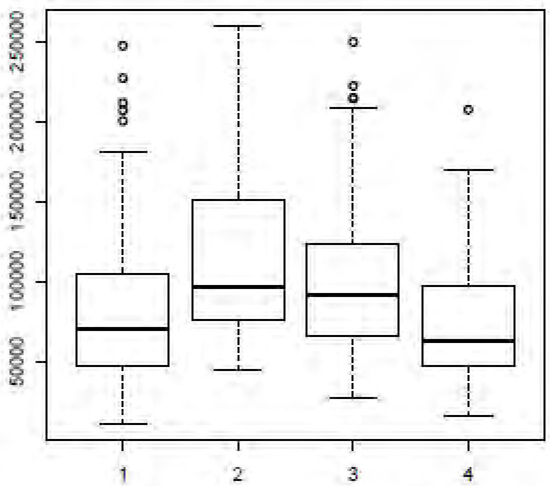


(c) Box plots for costs by region

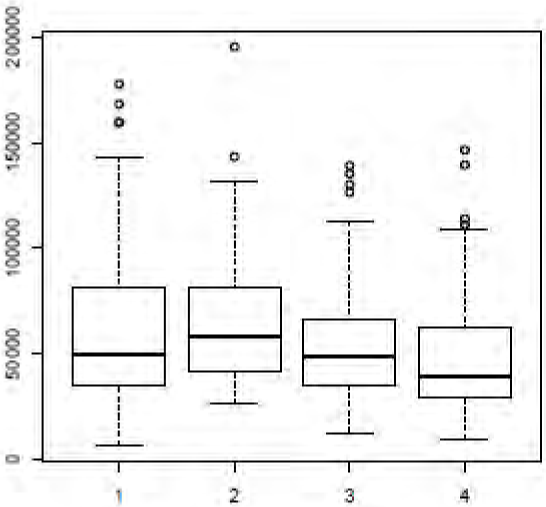
Bivariate - area, receipts, costs by region – SDC

		region			
		1	2	3	4
area	Up to 29	11	7	3	14
	30 – 39	18	7	16	29
	40 – 49	20	7	16	17
	50 – 59	18	6	9	10
	60 – 79	24	5	22	11
	80 and over	32	6	19	6

(a) area by region

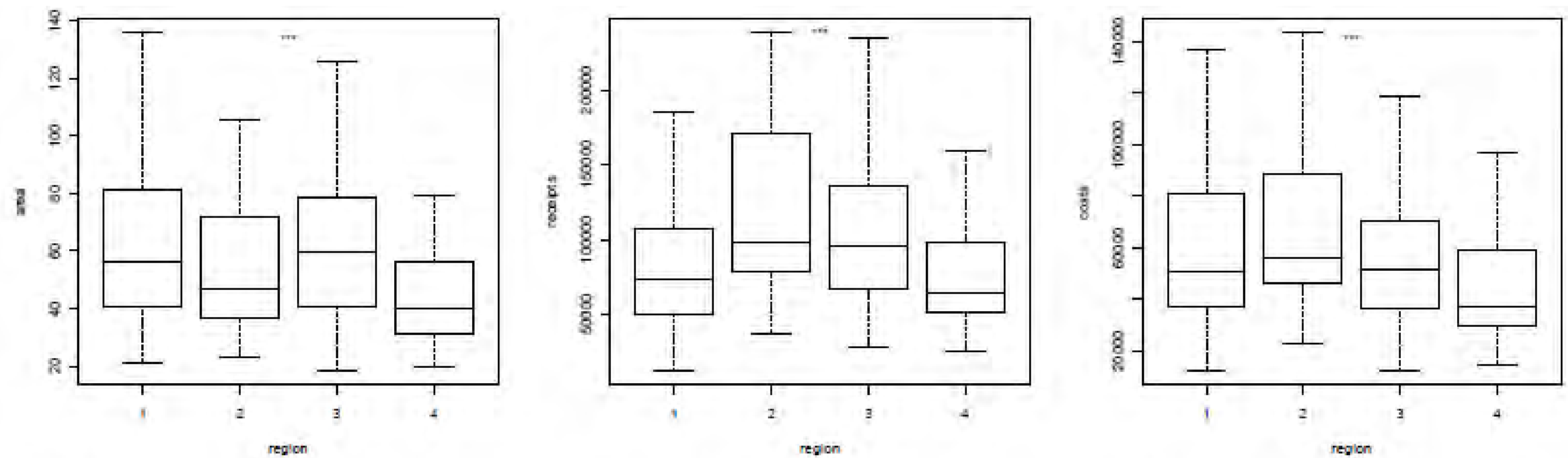


(b) receipts by region



(c) costs by region

Bivariate - area, receipts, costs by region – remote analysis

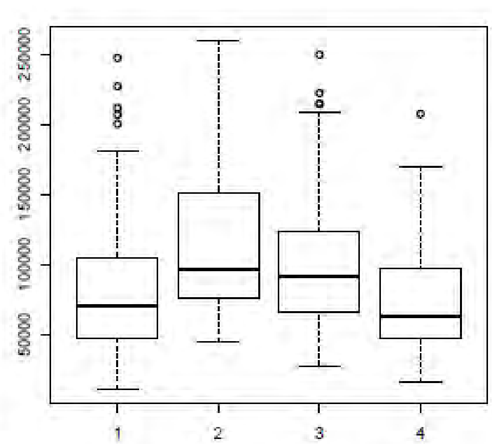


(a) Box plots for area by region (b) Box plots for receipts by region (c) Box plots for costs by region

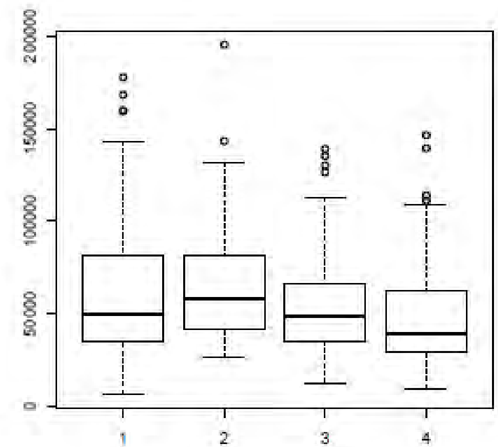
Bivariate - area, receipts, costs by region – side by side

		region			
		1	2	3	4
area	Up to 29	11	7	3	14
	30 – 39	18	7	16	29
	40 – 49	20	7	16	17
	50 – 59	18	6	9	10
	60 – 79	24	5	22	11
	80 and over	32	6	19	6

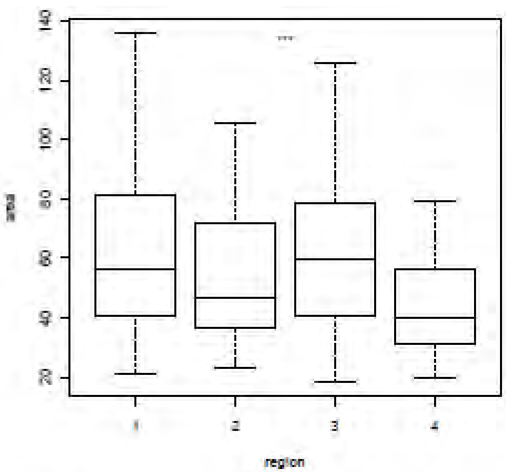
(a) area by region



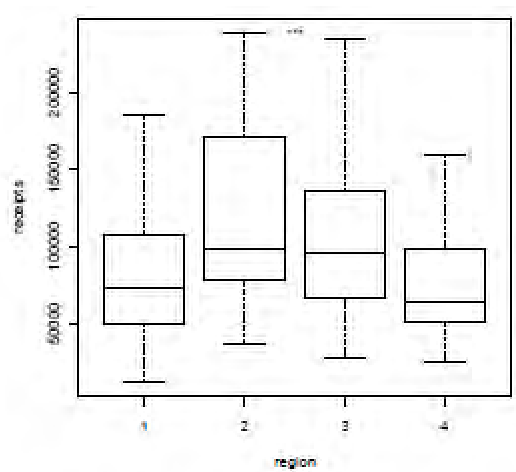
(b) receipts by region



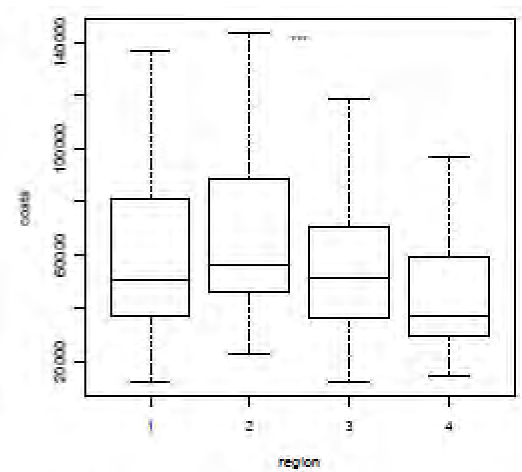
(c) costs by region



(a) Box plots for area by region

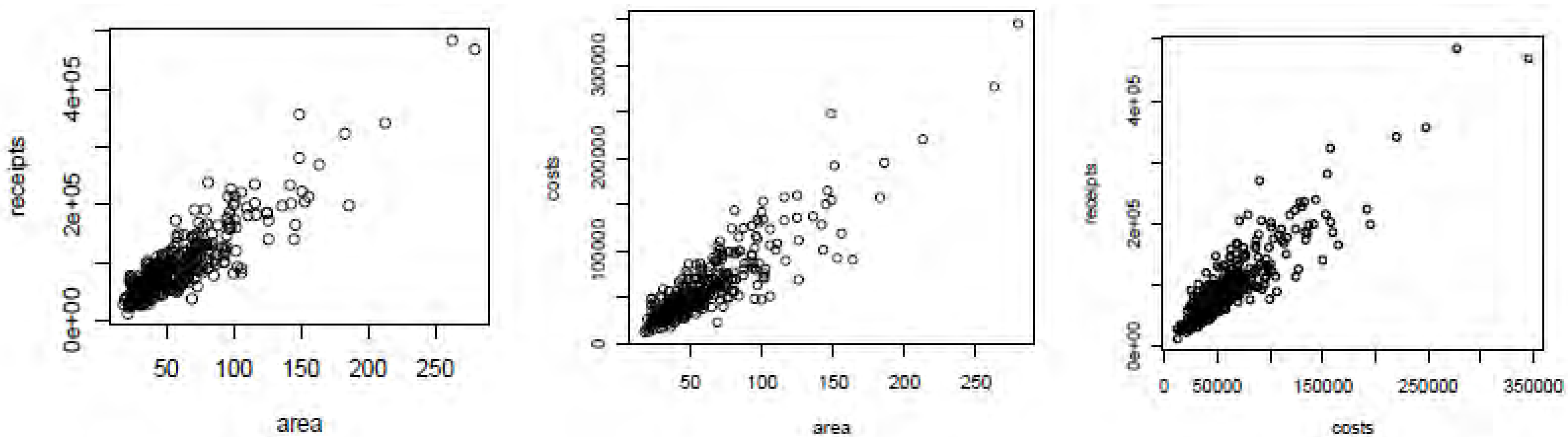


(b) Box plots for receipts by region



(c) Box plots for costs by region

Bivariate – pairs from area, receipts costs - unconfidentialised

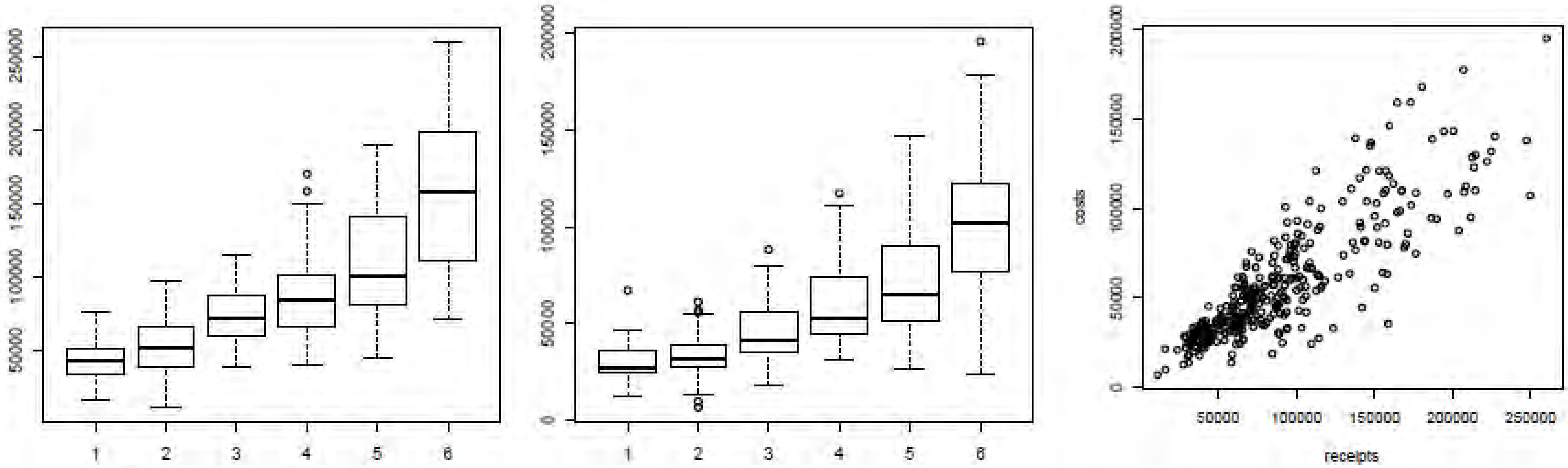


(a) receipts by area (b) costs by area (c) receipts by costs

	area	receipts	costs
area	1	0.8876671 p-value < 2.2e-16	0.8867933 p-value < 2.2e-16
receipts		1	0.90096490 p-value < 2.2e-16
costs			1

(d) Pearson Correlation Coefficients

Bivariate – pairs from area, receipts costs – SDC

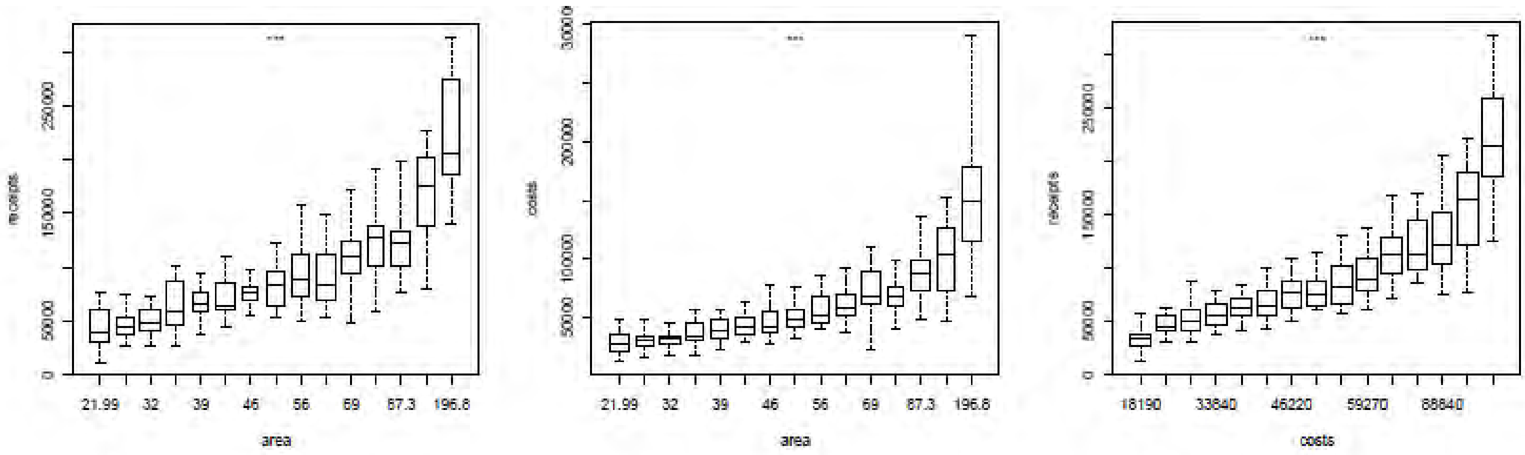


(a) receipts by area (b) costs by area (c) receipts by costs

	area	receipts	costs
area	1	0.7487469 p-value < 2.2e-16	0.7145667 p-value < 2.2e-16
receipts		1	0.8594960 p-value < 2.2e-16
costs			1

(d) Pearson Correlation Coefficients

Bivariate – pairs from area, receipts costs – remote analysis



(a) receipts vs area (b) costs vs area (c) receipts vs costs

	area	receipts	costs
area	1	0.8877 ***	0.8868 ***
receipts		1	0.9010 ***
costs			1

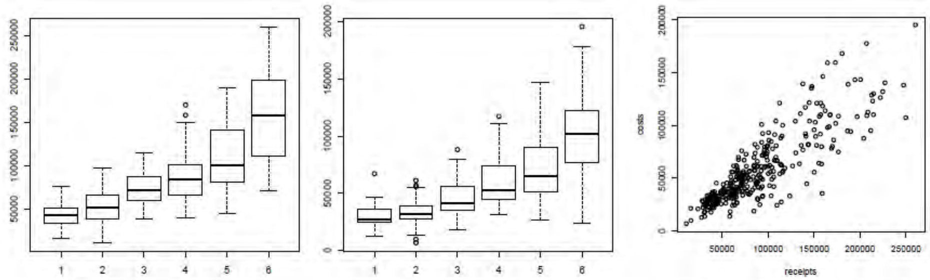
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(d) Pearson Correlation Coefficients

$$\chi^2 = 350 \qquad *** \qquad C.V. = 0.45$$

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

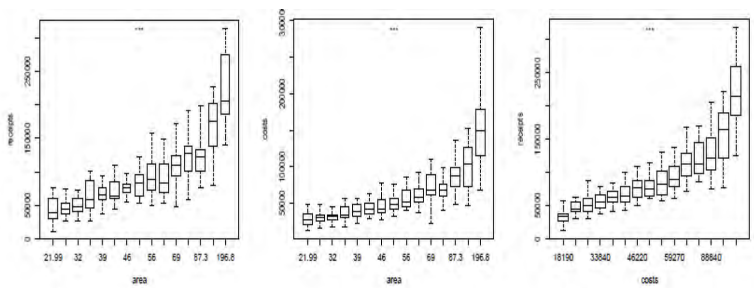
Bivariate – pairs from area, receipts costs – side by side



(a) receipts by area (b) costs by area (c) receipts by costs

	area	receipts	costs
area	1	0.7487469 p-value < 2.2e-16	0.7145667 p-value < 2.2e-16
receipts		1	0.8594960 p-value < 2.2e-16
costs			1

(d) Pearson Correlation Coefficients



(a) receipts vs area (b) costs vs area (c) receipts vs costs

	area	receipts	costs
area	1	0.8877 ***	0.8868 ***
receipts		1	0.9010 ***
costs			1

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(d) Pearson Correlation Coefficients

$\chi^2 = 350$ *** $C.V. = 0.45$

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Example

Linear Regression

Sugar Farms Data

- Response = log(receipts)
- Explanatory = region, area, log(harvest), log(costs)

- Confidentialised Input – SDC
  - Outliers are deleted
  - Area is categorised into 6 bands
  - Noise is added to receipts, costs, profit to preserve correlations
- Confidentialised Output – Remote Analysis
  - Confidentialisation filters applied to output
- Unconfidentialised
  - Traditional approach

Summary Results

	Confidentialised Input	Confidentialised Output	Un-confidentialised
Intercept p-value significance	3.627253 < 2e-16 ***	3.06 ***	2.7060226 < 2e-16 ***
Factor(region)2 p-value significance	0.192557 2.97e-15 ***	0.205 ***	0.1814301 < 2e-16 ***
Factor(region)3 p-value significance	0.187611 < 2e-16 ***	0.244 ***	0.2390758 < 2e-16 ***
Factor(region)4 p-value significance	0.091021 1.91e-7 ***	0.117 ***	0.1184681 < 2e-16 ***
area p-value significance	0.031205 4.81e-6 ***	0.0004	0.0000792 0.773
harvest p-value significance	0. 831541 < 2e-16 ***	0.883 ***	0.8655644 < 2e-16 ***
costs p-value signficance	0. 063136 0.0147 *	0.0823 ***	0.1309820 4.05e-8 ***

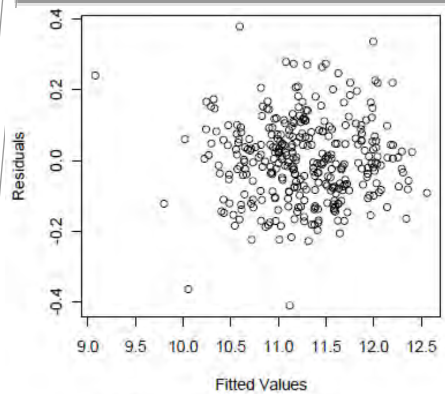
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Goodness of Fit statistics

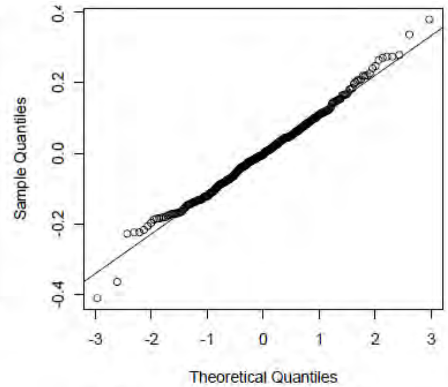
	Confidentialised Input	Confidentialised Output	Un-confidentialised
Residual standard error degrees of freedom	0.1151 326	0.08 314	0.09024 331
Multiple R squared	0.9554	0.97	0.974
Adjusted R squared	0.9546	0.97	0.9735
F-statistic degrees of freedom p-value significance	1164 6 and 326 < 2.2e-16 ***	2100 6 and 331 - ***	2067 6 and 331 < 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

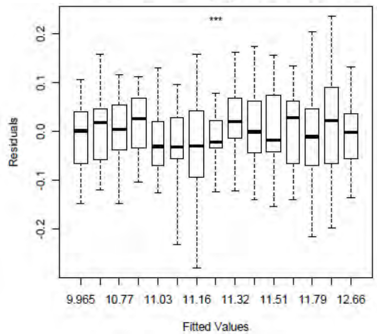
Model diagnostics



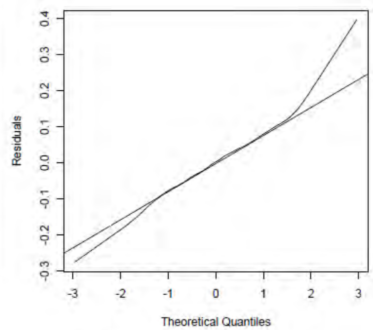
(a) Residuals vs Fitted Values



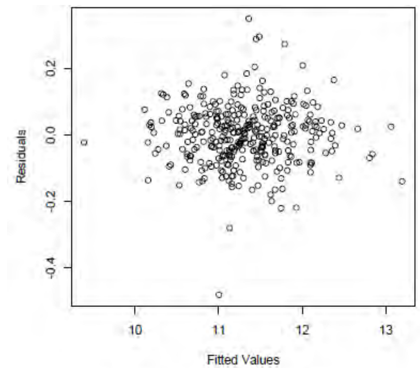
(b) Normal Q-Q Plot of Residuals



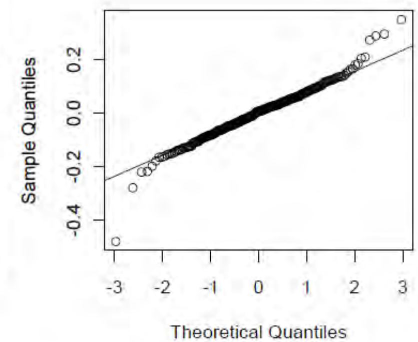
(a) Residuals by fitted values



(b) Normal QQ Plot of Residuals



(a) Residuals vs Fitted Values



(b) Normal Q-Q Plot of Residuals

Summary

Summary

- Remote Analysis & Differential Privacy
  - Logistic regression
  - Other models…

Headline conclusion:

can be better to add noise to something other than the output

- Remote Analysis vs Statistical Disclosure Control
  - Business Data

Headline conclusion:

remote analysis seems preferable

References

- K. Chaudhuri and C. Monteleoni, Privacy-preserving logistic regression, Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS), 2008, 289-296.
- A. Smith, Efficient, differentially private point estimators, 2008. ArXiv:0809.4794v1.

- R. Sparks, C. Carter, J.B. Donnelly, C.M. O'Keefe, J. Duncan, T. Keighley and D. McAullay, Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving AnalyticsTM, Comput Methods Programs Biomed 91 (2008), 208-222.
- C.M. O'Keefe and N.M. Good, Regression Output from a Remote Analysis Server, Data & Knowledge Engineering 68 (2009), 1175-1186.
- C.M. O'Keefe, Remote Analysis in Action – Design and Implementation of a Demonstration Remote Analysis System, Proceedings of the New Techniques and Technologies in Statistics conference NTTS 2011, Eurostat, Brussels 22-24 Feb 2011. Available at www.ntts2011.eu
- C.M. O'Keefe, Confidentialising exploratory data analysis output in a remote analysis system, submitted.
- C.M. O'Keefe and N. Shlomo, Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data, submitted.

CSIRO Mathematics, Informatics and Statistics
Prof Christine O'Keefe PhD MBA
Research Leader, Privacy and Confidentiality, CSIRO
Adjunct Professor, University of Adelaide

Phone: +61 2 6216 7021
Email: Christine.O'Keefe@csiro.au
Web: www.csiro.au/people/Christine.O'Keefe

www.csiro.au

Thank you

Contact Us
Phone: 1300 363 400 or +61 3 9545 2176
Email: enquiries@csiro.au Web: www.csiro.au