Hunting High and Low: Visualising Shifting Correlations in Financial Markets

Supporting Materials

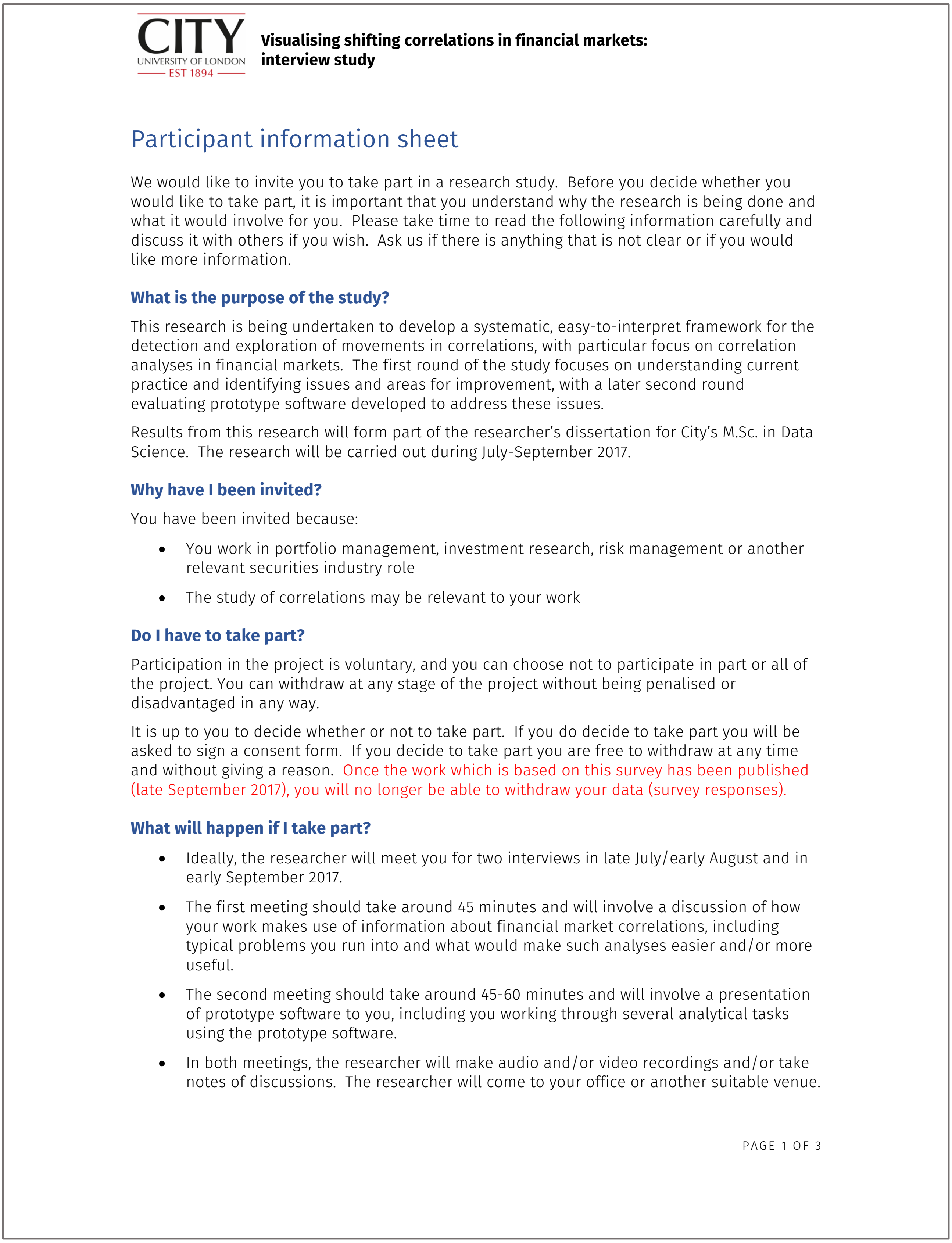
Contents

For further information please contact:

Peter M. Simon *peter@scaridae.com*
Dr. Cagatay Turkay *cagatay.turkay@city.ac.uk*

1 Supporting documents for user studies

1.1 Participant information sheet



CITY UNIVERSITY OF LONDON EST 1894

Visualising shifting correlations in financial markets: interview study

Participant information sheet

We would like to invite you to take part in a research study. Before you decide whether you would like to take part, it is important that you understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information.

What is the purpose of the study?

This research is being undertaken to develop a systematic, easy-to-interpret framework for the detection and exploration of movements in correlations, with particular focus on correlation analyses in financial markets. The first round of the study focuses on understanding current practice and identifying issues and areas for improvement, with a later second round evaluating prototype software developed to address these issues.

Results from this research will form part of the researcher's dissertation for City's M.Sc. in Data Science. The research will be carried out during July-September 2017.

Why have I been invited?

You have been invited because:

- You work in portfolio management, investment research, risk management or another relevant securities industry role
- The study of correlations may be relevant to your work

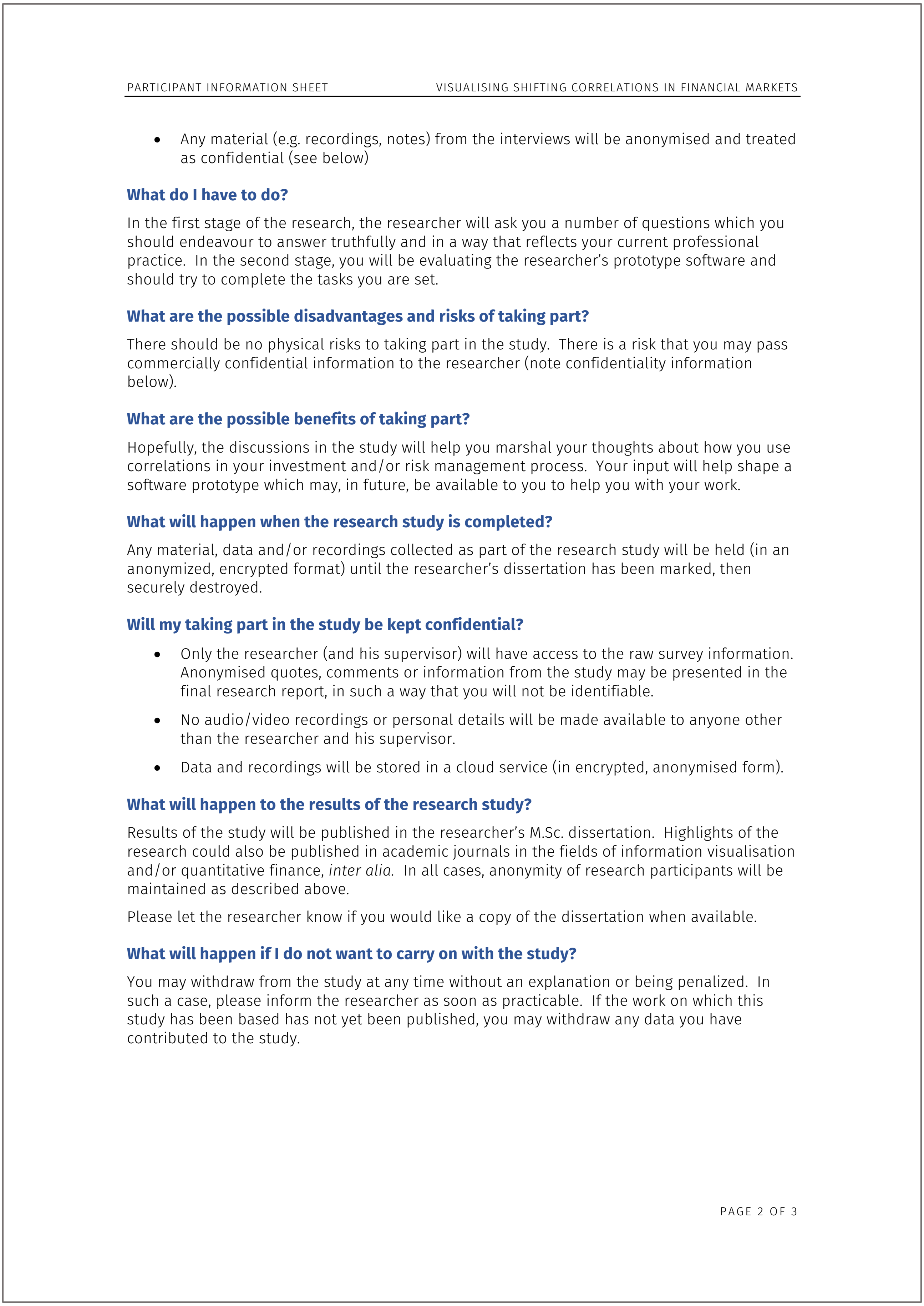
Do I have to take part?

Participation in the project is voluntary, and you can choose not to participate in part or all of the project. You can withdraw at any stage of the project without being penalised or disadvantaged in any way.

It is up to you to decide whether or not to take part. If you do decide to take part you will be asked to sign a consent form. If you decide to take part you are free to withdraw at any time and without giving a reason. Once the work which is based on this survey has been published (late September 2017), you will no longer be able to withdraw your data (survey responses).

What will happen if I take part?

- Ideally, the researcher will meet you for two interviews in late July/early August and in early September 2017.
- The first meeting should take around 45 minutes and will involve a discussion of how your work makes use of information about financial market correlations, including typical problems you run into and what would make such analyses easier and/or more useful.
- The second meeting should take around 45-60 minutes and will involve a presentation of prototype software to you, including you working through several analytical tasks using the prototype software.
- In both meetings, the researcher will make audio and/or video recordings and/or take notes of discussions. The researcher will come to your office or another suitable venue.



- Any material (e.g. recordings, notes) from the interviews will be anonymised and treated as confidential (see below)

What do I have to do?

In the first stage of the research, the researcher will ask you a number of questions which you should endeavour to answer truthfully and in a way that reflects your current professional practice. In the second stage, you will be evaluating the researcher's prototype software and should try to complete the tasks you are set.

What are the possible disadvantages and risks of taking part?

There should be no physical risks to taking part in the study. There is a risk that you may pass commercially confidential information to the researcher (note confidentiality information below).

What are the possible benefits of taking part?

Hopefully, the discussions in the study will help you marshal your thoughts about how you use correlations in your investment and/or risk management process. Your input will help shape a software prototype which may, in future, be available to you to help you with your work.

What will happen when the research study is completed?

Any material, data and/or recordings collected as part of the research study will be held (in an anonymized, encrypted format) until the researcher's dissertation has been marked, then securely destroyed.

Will my taking part in the study be kept confidential?

- Only the researcher (and his supervisor) will have access to the raw survey information. Anonymised quotes, comments or information from the study may be presented in the final research report, in such a way that you will not be identifiable.
- No audio/video recordings or personal details will be made available to anyone other than the researcher and his supervisor.
- Data and recordings will be stored in a cloud service (in encrypted, anonymised form).

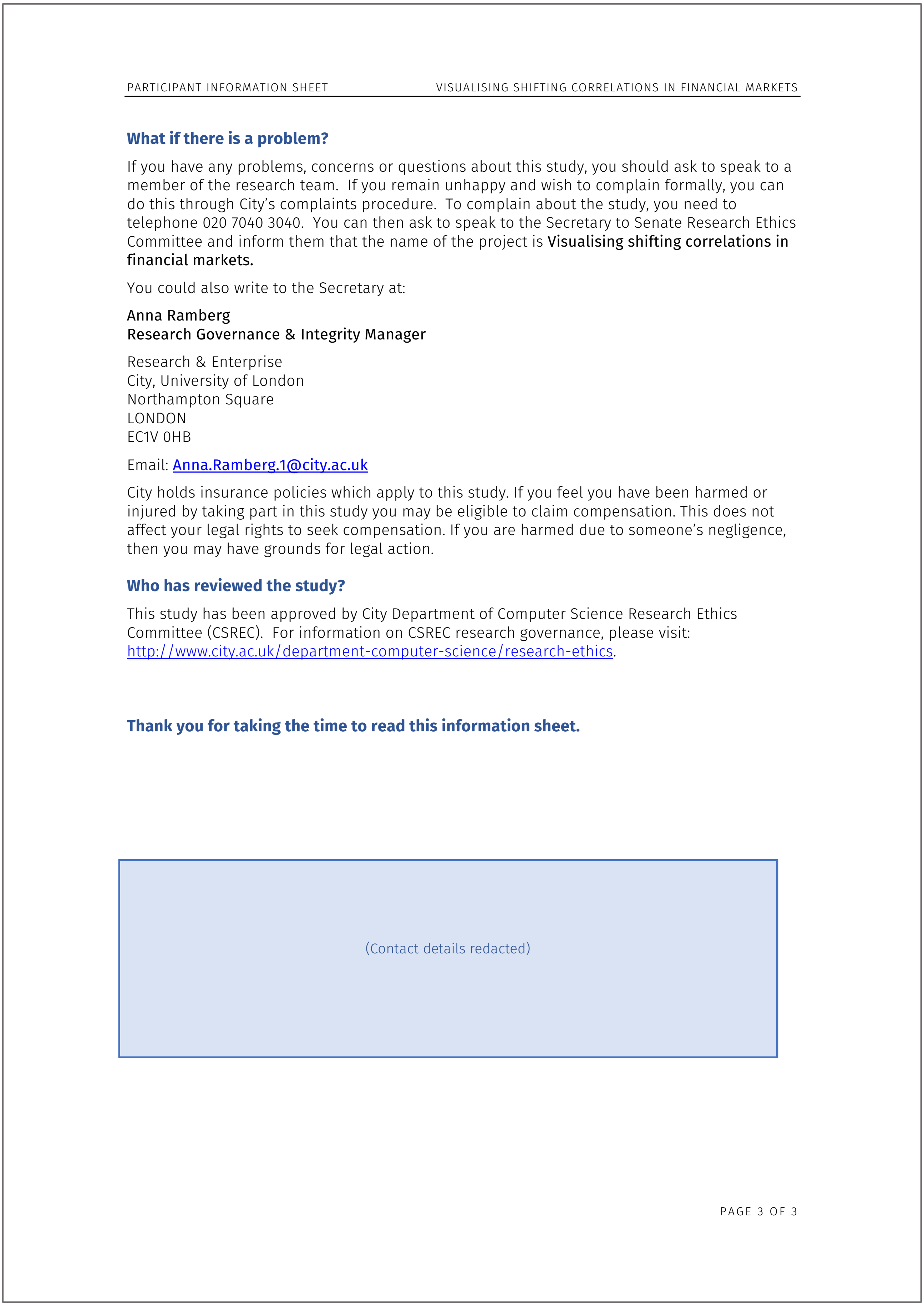
What will happen to the results of the research study?

Results of the study will be published in the researcher's M.Sc. dissertation. Highlights of the research could also be published in academic journals in the fields of information visualisation and/or quantitative finance, *inter alia*. In all cases, anonymity of research participants will be maintained as described above.

Please let the researcher know if you would like a copy of the dissertation when available.

What will happen if I do not want to carry on with the study?

You may withdraw from the study at any time without an explanation or being penalized. In such a case, please inform the researcher as soon as practicable. If the work on which this study has been based has not yet been published, you may withdraw any data you have contributed to the study.



What if there is a problem?

If you have any problems, concerns or questions about this study, you should ask to speak to a member of the research team. If you remain unhappy and wish to complain formally, you can do this through City's complaints procedure. To complain about the study, you need to telephone 020 7040 3040. You can then ask to speak to the Secretary to Senate Research Ethics Committee and inform them that the name of the project is **Visualising shifting correlations in financial markets.**

You could also write to the Secretary at:

Anna Ramberg
Research Governance & Integrity Manager

Research & Enterprise
City, University of London
Northampton Square
LONDON
EC1V 0HB

Email: Anna.Ramberg.1@city.ac.uk

City holds insurance policies which apply to this study. If you feel you have been harmed or injured by taking part in this study you may be eligible to claim compensation. This does not affect your legal rights to seek compensation. If you are harmed due to someone's negligence, then you may have grounds for legal action.

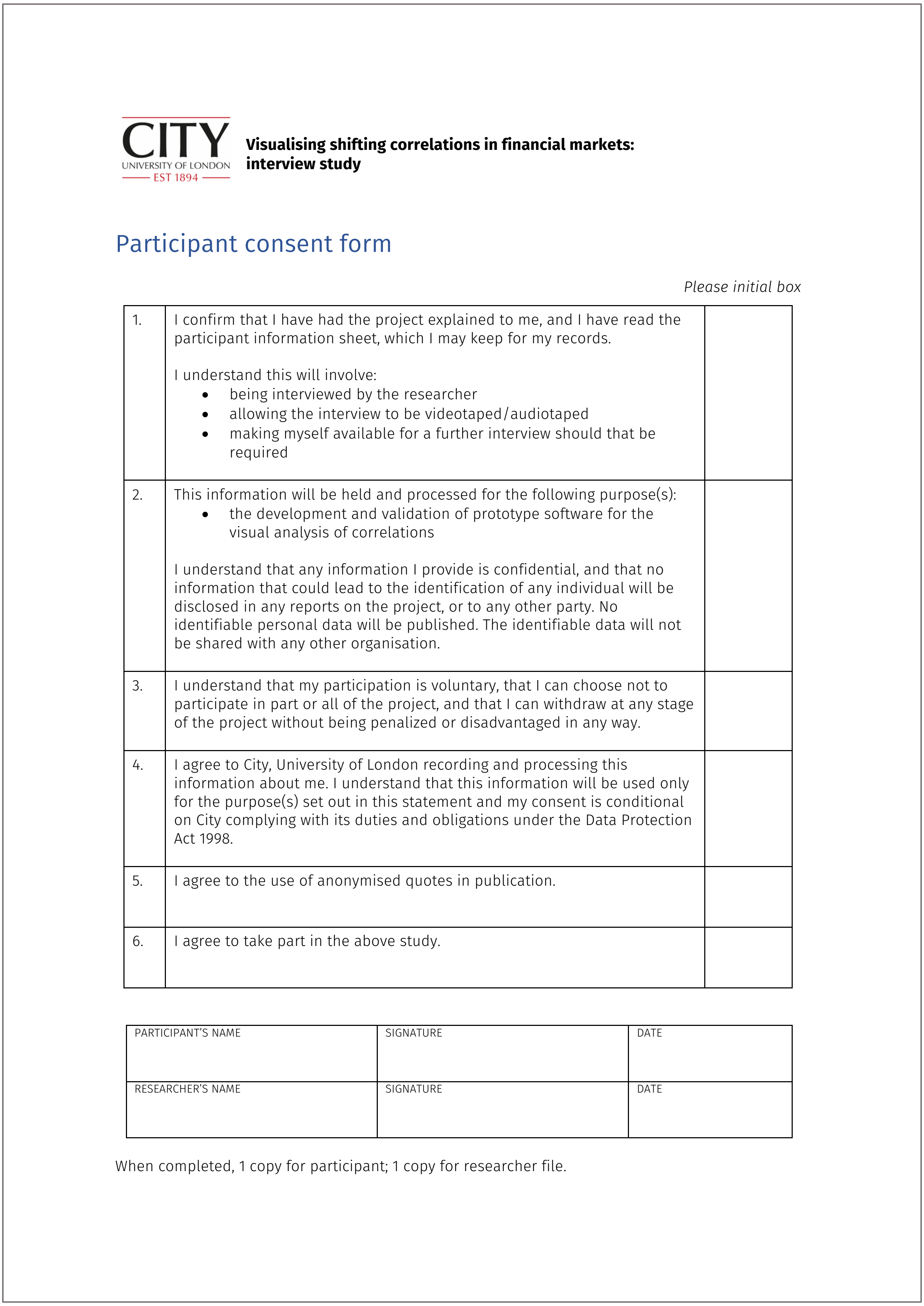
Who has reviewed the study?

This study has been approved by City Department of Computer Science Research Ethics Committee (CSREC). For information on CSREC research governance, please visit: http://www.city.ac.uk/department-computer-science/research-ethics.

Thank you for taking the time to read this information sheet.

(Contact details redacted)

1.2 Participant consent form



CITY
UNIVERSITY OF LONDON
EST 1894

Visualising shifting correlations in financial markets: interview study

Participant consent form

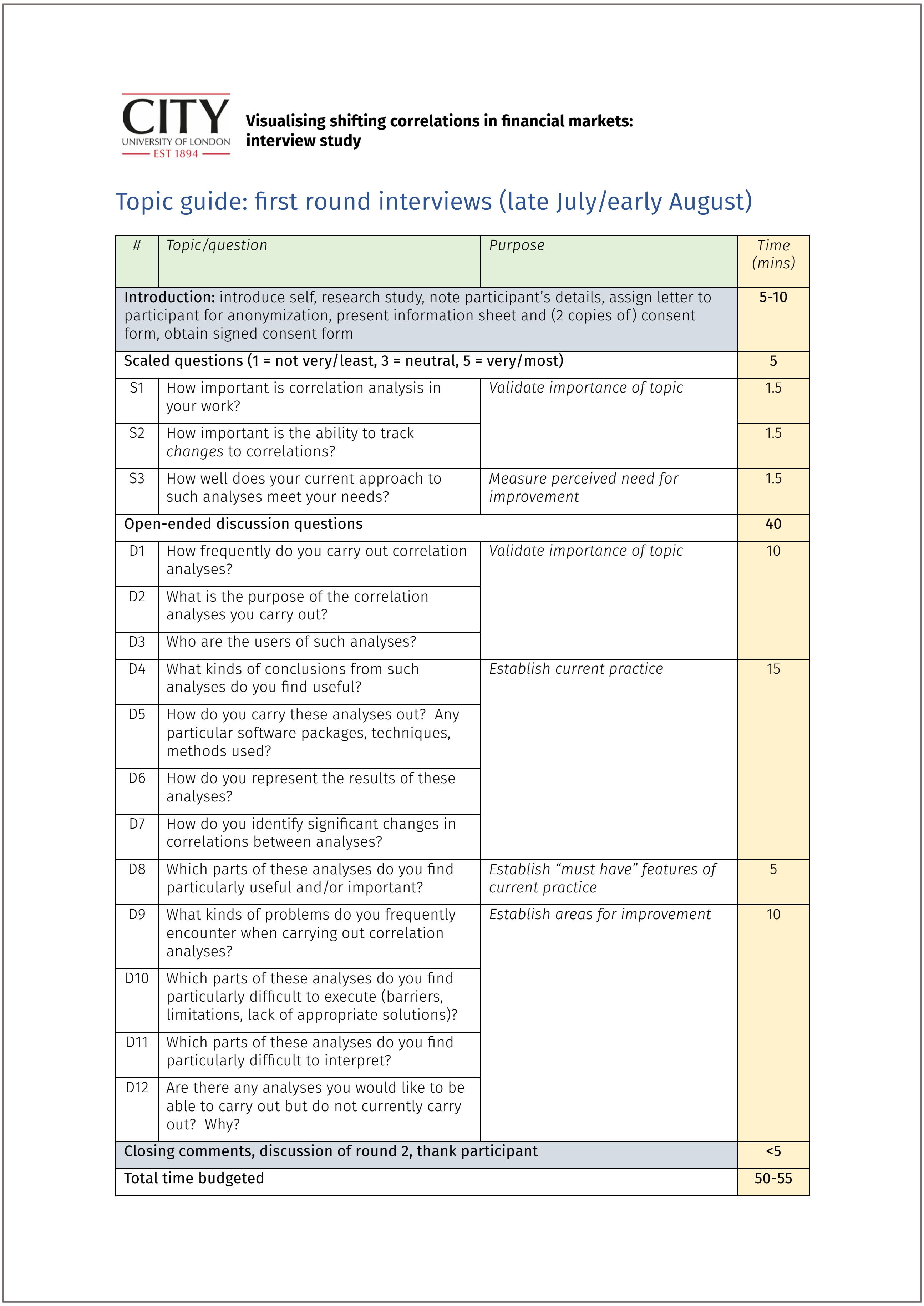
Please initial box

1.	I confirm that I have had the project explained to me, and I have read the participant information sheet, which I may keep for my records. I understand this will involve: • being interviewed by the researcher • allowing the interview to be videotaped/audiotaped • making myself available for a further interview should that be required	
2.	This information will be held and processed for the following purpose(s): • the development and validation of prototype software for the visual analysis of correlations I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be disclosed in any reports on the project, or to any other party. No identifiable personal data will be published. The identifiable data will not be shared with any other organisation.	
3.	I understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdraw at any stage of the project without being penalized or disadvantaged in any way.	
4.	I agree to City, University of London recording and processing this information about me. I understand that this information will be used only for the purpose(s) set out in this statement and my consent is conditional on City complying with its duties and obligations under the Data Protection Act 1998.	
5.	I agree to the use of anonymised quotes in publication.	
6.	I agree to take part in the above study.	

PARTICIPANT'S NAME	SIGNATURE	DATE
RESEARCHER'S NAME	SIGNATURE	DATE

When completed, 1 copy for participant; 1 copy for researcher file.

1.3 Topic guide for interviewer: first round



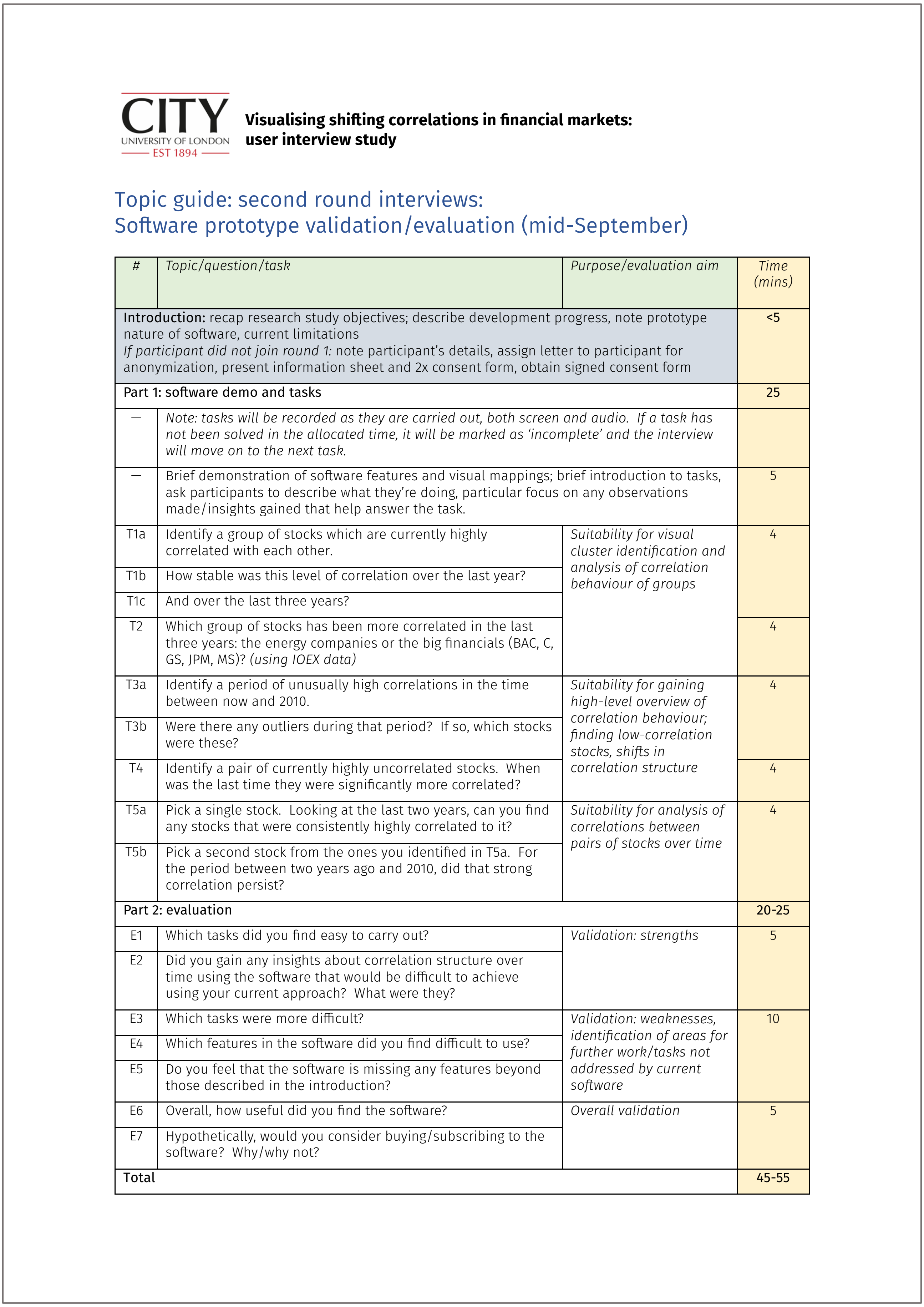
CITY
UNIVERSITY OF LONDON
EST 1894

Visualising shifting correlations in financial markets: interview study

Topic guide: first round interviews (late July/early August)

#	*Topic/question*	*Purpose*	*Time (mins)*
Introduction: introduce self, research study, note participant's details, assign letter to participant for anonymization, present information sheet and (2 copies of) consent form, obtain signed consent form			5-10
Scaled questions (1 = not very/least, 3 = neutral, 5 = very/most)			5
S1	How important is correlation analysis in your work?	*Validate importance of topic*	1.5
S2	How important is the ability to track *changes* to correlations?		1.5
S3	How well does your current approach to such analyses meet your needs?	*Measure perceived need for improvement*	1.5
Open-ended discussion questions			40
D1	How frequently do you carry out correlation analyses?	*Validate importance of topic*	10
D2	What is the purpose of the correlation analyses you carry out?		
D3	Who are the users of such analyses?		
D4	What kinds of conclusions from such analyses do you find useful?	*Establish current practice*	15
D5	How do you carry these analyses out? Any particular software packages, techniques, methods used?		
D6	How do you represent the results of these analyses?		
D7	How do you identify significant changes in correlations between analyses?		
D8	Which parts of these analyses do you find particularly useful and/or important?	*Establish "must have" features of current practice*	5
D9	What kinds of problems do you frequently encounter when carrying out correlation analyses?	*Establish areas for improvement*	10
D10	Which parts of these analyses do you find particularly difficult to execute (barriers, limitations, lack of appropriate solutions)?		
D11	Which parts of these analyses do you find particularly difficult to interpret?		
D12	Are there any analyses you would like to be able to carry out but do not currently carry out? Why?		
Closing comments, discussion of round 2, thank participant			<5
Total time budgeted			50-55

1.4 Topic guide for interviewer: second round



Visualising shifting correlations in financial markets: user interview study

Topic guide: second round interviews: Software prototype validation/evaluation (mid-September)

#	*Topic/question/task*	*Purpose/evaluation aim*	*Time (mins)*
Introduction: recap research study objectives; describe development progress, note prototype nature of software, current limitations *If participant did not join round 1:* note participant's details, assign letter to participant for anonymization, present information sheet and 2x consent form, obtain signed consent form			<5
Part 1: software demo and tasks			25
—	*Note: tasks will be recorded as they are carried out, both screen and audio. If a task has not been solved in the allocated time, it will be marked as 'incomplete' and the interview will move on to the next task.*		
—	Brief demonstration of software features and visual mappings; brief introduction to tasks, ask participants to describe what they're doing, particular focus on any observations made/insights gained that help answer the task.		5
T1a	Identify a group of stocks which are currently highly correlated with each other.	*Suitability for visual cluster identification and analysis of correlation behaviour of groups*	4
T1b	How stable was this level of correlation over the last year?		
T1c	And over the last three years?		
T2	Which group of stocks has been more correlated in the last three years: the energy companies or the big financials (BAC, C, GS, JPM, MS)? *(using IOEX data)*		4
T3a	Identify a period of unusually high correlations in the time between now and 2010.	*Suitability for gaining high-level overview of correlation behaviour; finding low-correlation stocks, shifts in correlation structure*	4
T3b	Were there any outliers during that period? If so, which stocks were these?		
T4	Identify a pair of currently highly uncorrelated stocks. When was the last time they were significantly more correlated?		4
T5a	Pick a single stock. Looking at the last two years, can you find any stocks that were consistently highly correlated to it?	*Suitability for analysis of correlations between pairs of stocks over time*	4
T5b	Pick a second stock from the ones you identified in T5a. For the period between two years ago and 2010, did that strong correlation persist?		
Part 2: evaluation			20-25
E1	Which tasks did you find easy to carry out?	*Validation: strengths*	5
E2	Did you gain any insights about correlation structure over time using the software that would be difficult to achieve using your current approach? What were they?		
E3	Which tasks were more difficult?	*Validation: weaknesses, identification of areas for further work/tasks not addressed by current software*	10
E4	Which features in the software did you find difficult to use?		
E5	Do you feel that the software is missing any features beyond those described in the introduction?		
E6	Overall, how useful did you find the software?	*Overall validation*	5
E7	Hypothetically, would you consider buying/subscribing to the software? Why/why not?		
Total			45-55

2 Round 1 (requirements gathering) user study: detailed results

2.1 Survey period and respondents

First-round interviews were conducted between 2 and 9 August 2017 (inclusive).

Six individuals were interviewed in five interviews:

- The head of risk and a member of the risk management team at a medium-size diversified hedge fund manager (**P1**; interviewed together),
- a senior quantitative analyst (and former risk manager) at a specialist global advisory firm for currency risk management (**P2**),
- the chief operating officer of a small start-up sector hedge fund (**P3**), previously working as risk manager of another hedge fund,
- the head (and lead portfolio manager) of a quantitative investment strategies group (**P4**), and
- a senior quantitative analyst at a large independent global asset management firm (**P5**).

Disclosure of personal relationships with the researcher:
***P3** and (one of) **P1** are former colleagues of the researcher; **P2** and **P5** are part-time M.Sc. Data Science students at City; **P4** is a part-time Ph.D. student at City.*

2.2 Thematic analysis

Table 2-a overleaf shows the themes that emerged from the interviews.

Table 2-a
Thematic analysis of requirements gathering survey results

■ *= mentioned* □ *= discussed; see footnote*

		Respondents				
		P1	P2	P3	P4	P5
	Scaled questions *(1=least/not very; 3=neutral; 5=most/very)*					
S1	How important is correlation analysis in your work?	**5**	**3-4**[1]	**4**	**4**[2]	**4**
S2	How important is the ability to track changes to correlations?	**5**	**3-5**[1]	**5**	**4**[2]	**4**
S3	How well does your current approach to such analyses meet your needs?	**3-4**	**1-2**	**3**	**3**	**3-4**
	Open-ended (discussion) questions					
D1	How frequently do you carry out correlation analyses?					
	Daily	■		■		■
	Varies/ad-hoc	■	■			■
	Infrequently (every 6-12 months)				■	
D2	What is the purpose of the correlation analyses you carry out?					
	Risk analysis	■	■	■	■	■
	Portfolio construction/analysis/risk budgeting		■	■	■	■
	Investment screening/idea generation		■			■
D3	Who are the users of such analyses?					
	Portfolio managers/investment team	■	■	■	■	■
	Risk management/regulatory team	■	■	■	■	■
	Firm management	■		■		
	Clients	■	■		□[3]	
D4	What kinds of conclusions from such analyses do you find useful?					
	Evidence of portfolio skews, unintended factors	■				
	Evidence of previously unknown risks	■				
	Factor-based analyses	■				
	Scenario analyses	■		■		
	Discovery of diversifiers/uncorrelated assets		■		■	■
	Discovery of groups of highly correlated assets/buckets				■	■
	Ex-post risk analyses/explanation of drawdowns			■		

(continues overleaf)

[1] Dependent on market environment; more important in times of market stress

[2] Implicit in work rather than explicit

[3] Would like to have good representation for client use, currently don't have

■ = *mentioned* □ = *discussed; see footnote*		*Respondents*				
		P1	P2	P3	P4	P5
Open-ended (discussion) questions *(continued)*						
D5	How do you carry these analyses out? Software packages, techniques, methods used?					
	APT risk management system	■				
	Bloomberg *PORT* (risk management functionality)			■		
	Microsoft Excel		■	■		
	Pertrac (hedge fund analysis software)		■			
	RiskMetrics risk management system			■		■
	Proprietary software (developed in-house)				■	■
	Data sources:					
	Bloomberg	■		■	■	■
	Datastream		■			
	Factset				■	■
	FE (Financial Express, funds database)		■			
D6	How do you represent the results of these analyses?					
	Tabular correlation matrices	■	■	□[4]		
	Heatmapped correlation matrices	■	■			■
	Narrative	■				
	Line charts of rolling correlations	■	■	■	■	
	Histograms (frequency of buckets in correl. matrix)	■				
	Scatter plots	■		■		
	Cluster analyses	■				■
	Range charts	■	■			
	Task and audience dependent	■				
	Standardised reports			■		
	Network graphs: minimum spanning trees					■
	Generally do not visualise results as implicit to work				■	
D7	How do you identify significant changes in correlations between analyses?					
	Graphically	■		■	■	■
	Difficult to do	■				
	Manual monitoring/meetings (compare matrices)	■	■	■		■
	Subtract correlation matrices for two points in time		■			
	Time series analysis/line charts of pairwise correlations	■	■	■	■	
D8	Which parts of these analyses do you find particularly useful and/or important?					
	Scenario and stress tests	■			■	
	Factor development	■				
	Ex-ante measures	■		■		
	Ex-post measures/simulation of past portfolio returns			■	■	
	Comparison between ex-ante and ex-post metrics			■		
	Simple representations for fund managers	■	■			
	In-depth analyses for risk managers	■				
	Identification of changes in correlations		■	□[5]	■	
	Identification of clusters/themes		■		□[6]	■
	Significance testing					■

(continues overleaf)

[4] Doesn't find this type of representation very useful, but produces for regulatory requirements

[5] "Especially if there is no change in the portfolio's composition"

[6] Particularly clusters with high correlation within and low correlation to other clusters

■ = *mentioned* ☐ = *discussed; see footnote*

		Respondents				
		P1	P2	P3	P4	P5
Open-ended (discussion) questions *(continued)*						
D9	**What kinds of problems do you frequently encounter when carrying out correlation analyses?**					
	Missing data points or series	■		■		
	Low and/or statistically insignificant correlations	■				■
	Changes in asset behaviour	■				
	Missing context	■				
	Too much information/“getting lost in numbers”		■			■
	Visual clutter		■			■
	Difficult to get meaningful output		■			■
	Structural differences in correlation (some assets more highly correlated than others)		■			
	Analyses less meaningful/accurate in times of stress			■		
	Numbers meaningless without context			■		
	“Much easier these days”			■		
	Assumption of parametric distribution of returns				■	
	Identification of securities with stable correlations				■	
	Asymmetry of correlations/variation by market env’t				■	
D10	**Which parts of these analyses do you find particularly difficult to interpret?**					
	Don’t know	■				
	Explaining correlation change/finding causality	■				
	Multi-factor correlations	■				
	Changes in correlations over time		■	■		
	Complexity		■			■
	Changes in cluster/bucket/category		■			
	Creating easy-to-read visualisation for client use				■	
D11	**Are there any analyses you would like to be able to carry out but do not currently carry out? Why?**					
	Daily tracking of funds’ correlations	■				
	Seeing evidence of crowding/clustering	■	■		■	
	Generally happy with what has been achieved		■			
	Tie in other portfolio characteristics			■		■
	Tie in with portfolio attribution			■		
	Better risk budgeting				■	
	Better understanding of asymmetric dynamics of risk				■	
Themes from concluding informal discussion: ideal attributes for software for visualising correlations						
	Ability to drill down	■				
	Ease of use	■			■	■
	Responsiveness	■		■		
	Flexibility			■	■	■
	Ability to export data	■				
	Easy-to-interpret graphical representation	■	■		■	■
	Ability to dynamically change time period & frequency	■				
	Ability to add in factors	■				
	Simplicity of representation		■			
	Highlight important conclusions		■			
	Clean data, no survivorship bias			■		
	Transparent methodology			■		
Miscellaneous						
	Open to participation in evaluation round of research	■	■	■	■	■
	Request for copy of finished work	■	■		■	■

3 Round 2 (software evaluation) study: detailed results

Table 3-a
Theme analysis of evaluation study

■ = *mentioned* □ = *somewhat*

		Respondents				
		P1	P2	P3	P4	P5
Part 1: task completion times (min:sec) — *to nearest 10 seconds* *4 minutes allocated per task/task group, marked incomplete (×) if not completed by then*						
T1a	Identify a group of stocks which are currently highly correlated with each other.	**0:10**	**1:00**	**0:10**	**0:30**	**0:10**
T1b	How stable was this level of correlation over the last year?	**0:20**	**0:20**	**1:30**	**0:20**	**0:30**
T1c	And over the last three years?	**0:50**	**0:20**	**0:20**	**0:10**	**0:30**
T2	Which group of stocks has been more correlated in the last three years: the energy companies or the big financials (BAC, C, GS, JPM, MS)?	**1:00**	**2:50**	**1:40**	**1:40**	**2:00**
T3a	Identify a period of unusually high correlations in the time between now and 2010.	**0:20**	**2:20**	**0:50**	**0:30**	**0:30**
T3b	Were there any outliers during that period? If so, which stocks were these?	**0:30**	×	**1:20**	**0:30**	**0:20**
T4	Identify a pair of currently highly uncorrelated stocks. When was the last time they were significantly more correlated?	**1:00**	**2:00**	**0:40**	**1:00**	**1:00**
T5a	Pick a single stock. Looking at the last two years, can you find any stocks that were consistently highly correlated to it?	**0:30**	**2:00**	**1:15**	**0:45**	**0:30**
T5b	Pick a second stock from the ones you identified in T5a. For the period between two years ago and 2010, did that strong correlation persist?	**0:20**	**1:00**	**0:20**	**0:30**	**0:30**
Part 2: evaluation questions—thematic analysis						
E1	Which tasks did you find easy to carry out?					
	T1 (identification of current high correlation)	■	■	■		
	T2 (comparison of two groups)		■	■		
	All			□[7]	■	■
E2	Did you gain any insights about correlation structure over time using the software that would be difficult to achieve using your current approach? What were they?					
	Illustration of correlation behaviour during market stress	□[8]	■	□[8]	■	
	The very high level of correlation amongst financials, so stock picking does not add value in that sector	■				
	Unable to see correlation dynamics using current approach		■			
	Persistency of correlation structure					■
	"Consistent with what I knew intuitively"				■	
	Similar insights gained as with current approach but much quicker and more intuitive to use than current system			■		
E3	Which tasks were more difficult?					
	None/all easy			■	■	■
	T2 (comparison of two groups)					□
	T3b (outlier identification)		■			
	T4 (identification/analysis of low correlation pair)		■		□	
	T5a (identification of consistently highly correlated stocks)	■				

(continues overleaf)

[7] All tasks seen as easy, T1 and T2 were easiest

[8] Implicit in reaction whilst watching an animation through 2011's correlation spike

■ = *mentioned* □ = *somewhat*

		P1	P2	P3	P4	P5
				Respondents		
Part 2: evaluation questions *(continued)*						
E4	**Which features in the software did you find difficult to use?**					
	None, very intuitive, only difficulties due to first-time use of software	■		■	■	
	Labelling needs improvement		■			■
	Animation — 'hard to get to grips with'		■			
	Zooming in on periods in line plots		■			
	Legends — too small, somewhat illegible		■			
	Absence of titles on line plots		■			
	Explanation by researcher poor, more detail needed		■			
	Text search to highlight stocks would be useful		■			
	Absence of help buttons/context help or similar					■
E5	**Do you feel that the software is missing any features beyond those described in the introduction?**					
	Easily change calculation period/frequency/data set	■	■	■	■	■
	Better time navigation (e.g. interact with line display to jump, linked crosshair across the three line charts)	■	■	■	□[9]	■
	Ability to visualise different types of pairwise data	■		■	■	□[10]
	More classification types, visual encodings (e.g. country)				■	■
	Correlation dispersion panel should be bigger, have more features (e.g. drill down)	■		■		
	Software should lead/guide user:		■			■
	Automatically highlight highly correlated links		■			
	Automatic outlier detection					■
	Automatic community detection/clustering				■	■
	Relate correlation dispersion panel and prices more closely, plot median correlation against prices, establish what predictive value/in what situations	■				
	Add display of index and/or stock volatility (either as circle size or as separate line display)	■				
	Numerical display in main plot		■			
	Histogram less useful for most users, should be optional		■			
	Load larger numbers of stocks into visualisation—"a few thousand"—and drill down into interactively				■	
	More filters on different criteria				■	
	Availability of graph theoretical measures					■
	More statistically robust calculation of correlations					■
	Ability to select, compare, plot two different groups					■
	Generally like features provided, shouldn't overcomplicate			■		
E6	**Overall, how useful did you find the software?**					
	Visually appealing		■			□[10]
	Straightforward to use		■	■		
	Liked visualisation of stocks		■	■		
	Liked *Sorter* view		■			
	Good for observing dynamics of correlations		■	■		
	A good start		■			
	Very good			■	■	■
	"7 or 8 out of 10"	■	■	■		■
	Innovative, very useful	■				

(continues overleaf)

[9] Implicit from participant's actions with mouse (attempt to click on line plot) during part 1

[10] Comment from participant's colleague

■ = *mentioned* □ = *somewhat*

		Respondents			
	P1	P2	P3	P4	P5
Part 2: evaluation questions *(continued)*					
E7 Hypothetically, would you consider buying/subscribing to the software? Why/why not?					
Yes definitely, subject to implementation of additional features discussed — “definitely would help me”	□[11]	■			
Yes, as is right now, if not very expensive					■
More likely to buy if had more features		■			■
Yes, would definitely consider it, price dependent	□[11]		■	■	
General reactions/comments					
“Never seen anything like it, innovative and very useful”	■				
“Would be very useful with portfolio data loaded, great way of [objectively] showing fund managers how effective or ineffective they are at stock picking”	■				
“Great tool for educating fund managers about diversification”	■				
“Innovative, novel, really good”	■				
“Useful for heavy users of correlations”		■			
“Adroit visualisation”		■			
“Cannot think of a way of capturing dynamics of correlations as powerfully as you have”		■			
“Fun to look at, I learned something”		■			
“Good work, so much better than looking at correlation matrices”			■		
“Really cool”, “fascinating”, “really impressive”, “awesome job”				■	
“Fascinating visualisation”				■	
“Very beautiful, great visualisation”					□[10]
“More or less what we do internally, but way better”					■
Colleagues joined to look at software	■				■

[11] “100% would buy if you can demonstrate predictive power for prices. If not, it’s still pretty useful, would certainly consider it, 50/50.” (A discussion of possible applications ensued.)

4 Final prototype performance statistics

Figure 4-1
Layout quality is better in times of market stress (high volatility), when distances are smaller

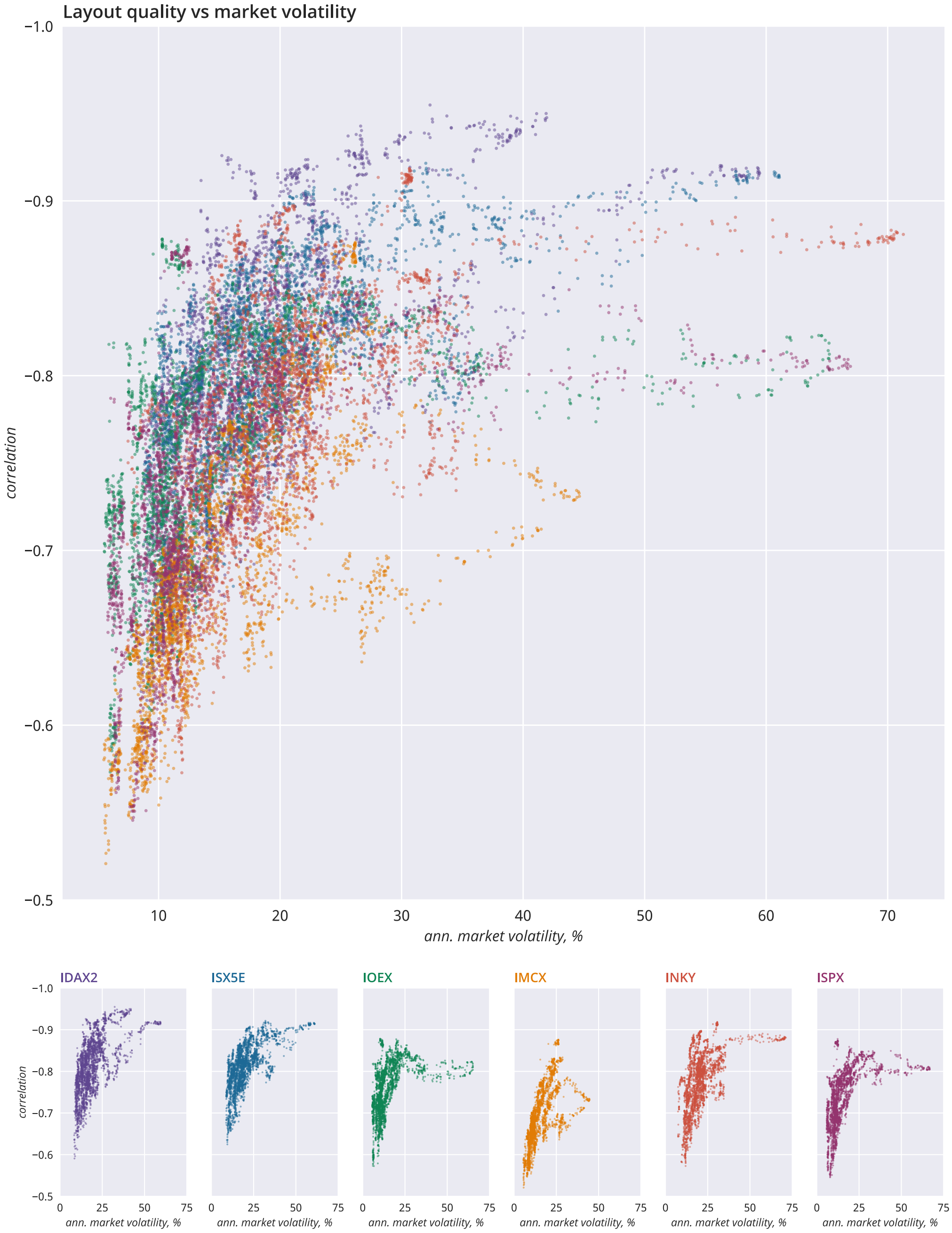
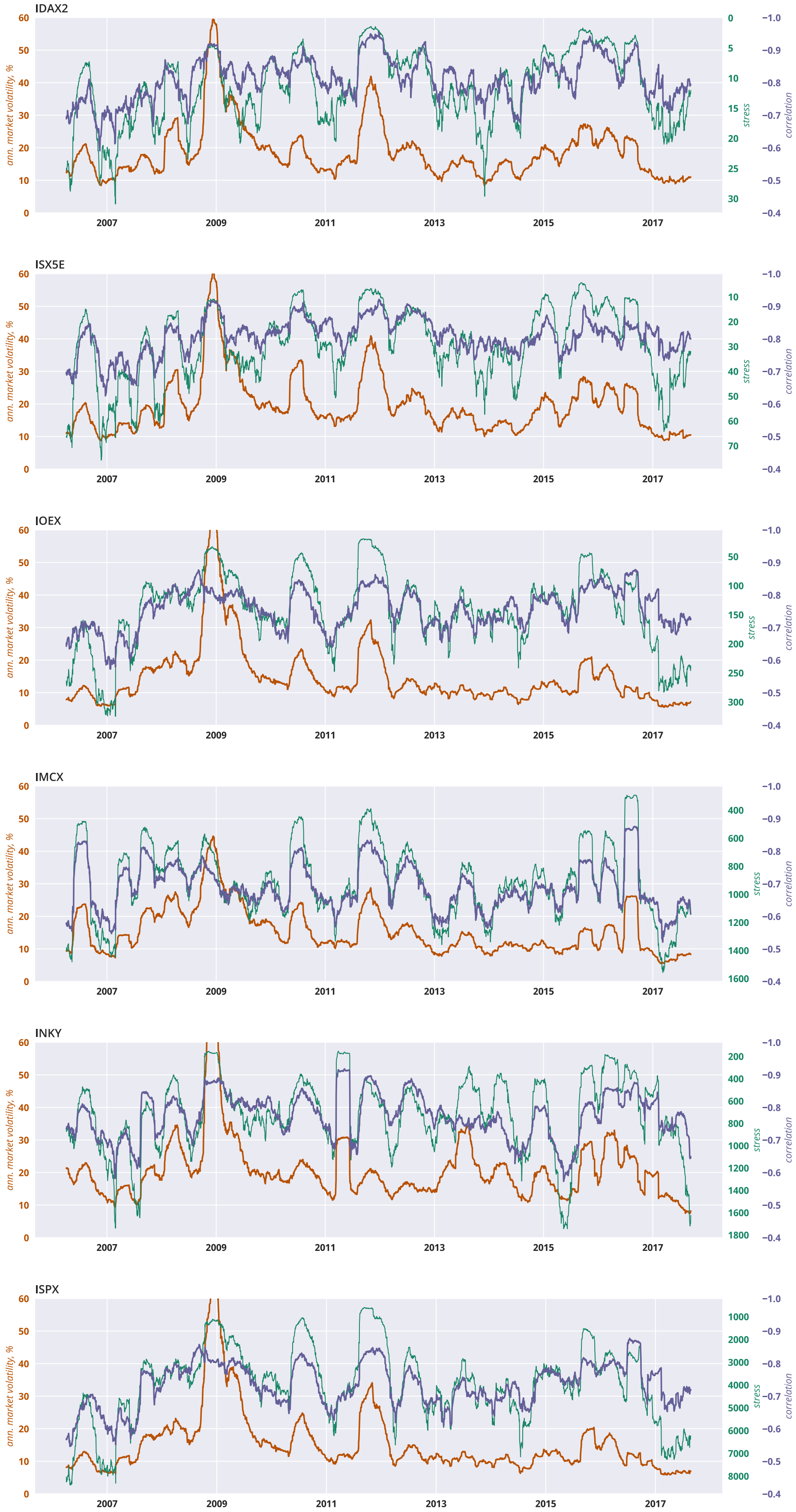


Figure 4-2

Market volatility, layout stress and layout quality* over time for the six test datasets*

**inverted scales*



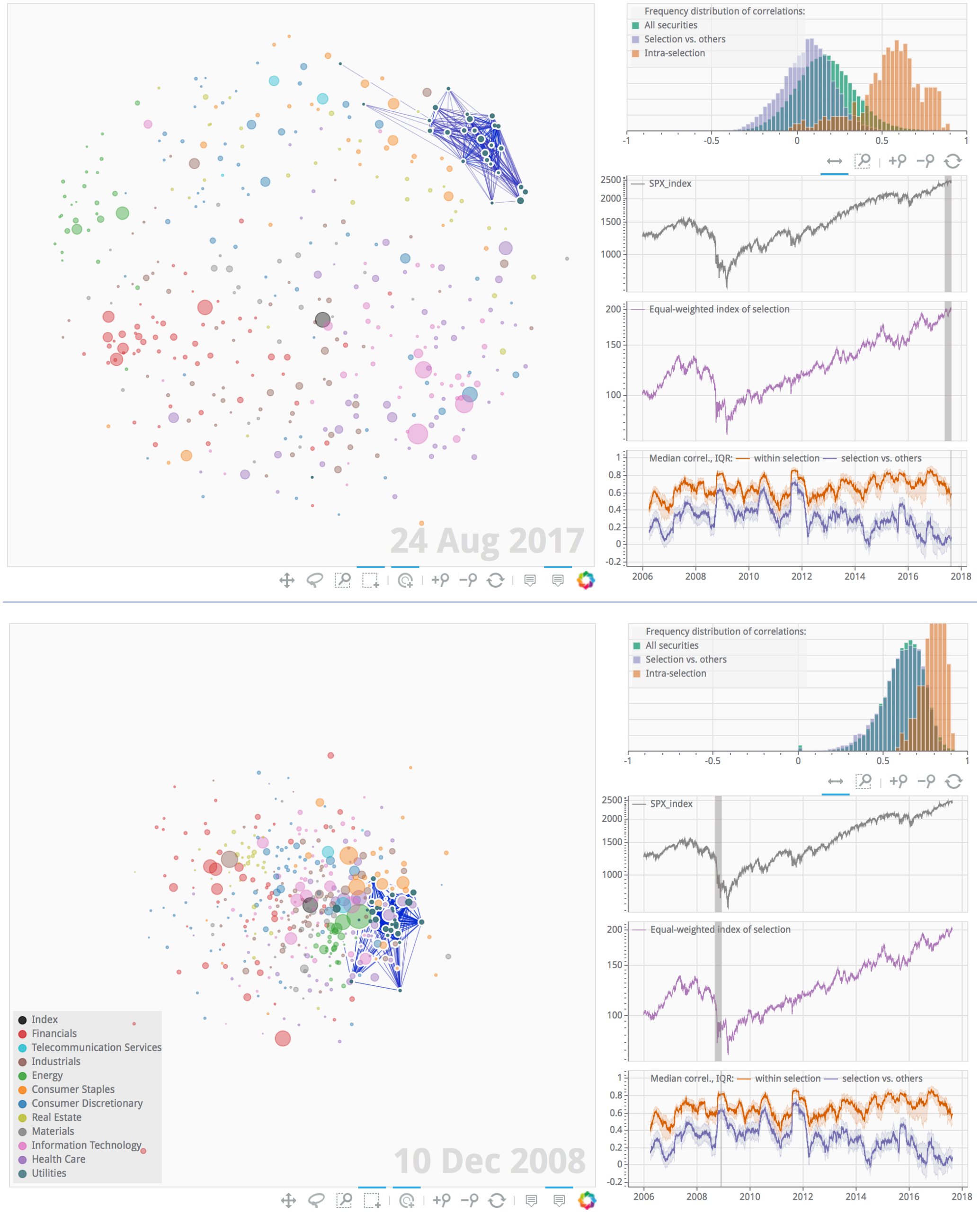
5 Further analysis stories

5.1.1 Visual cluster identification and correlation behaviour: the highly-correlated US utilities sector

Figure 5-1

c|swarm display for S&P 500 index with utilities sector highlighted

Even in times of market stress (the global financial crisis of 2008, lower panel) and generally high correlations, the group clusters closely together, with stocks within the sector being much more correlated to each other than to stocks in other sectors; note the histograms and the median correlations line charts

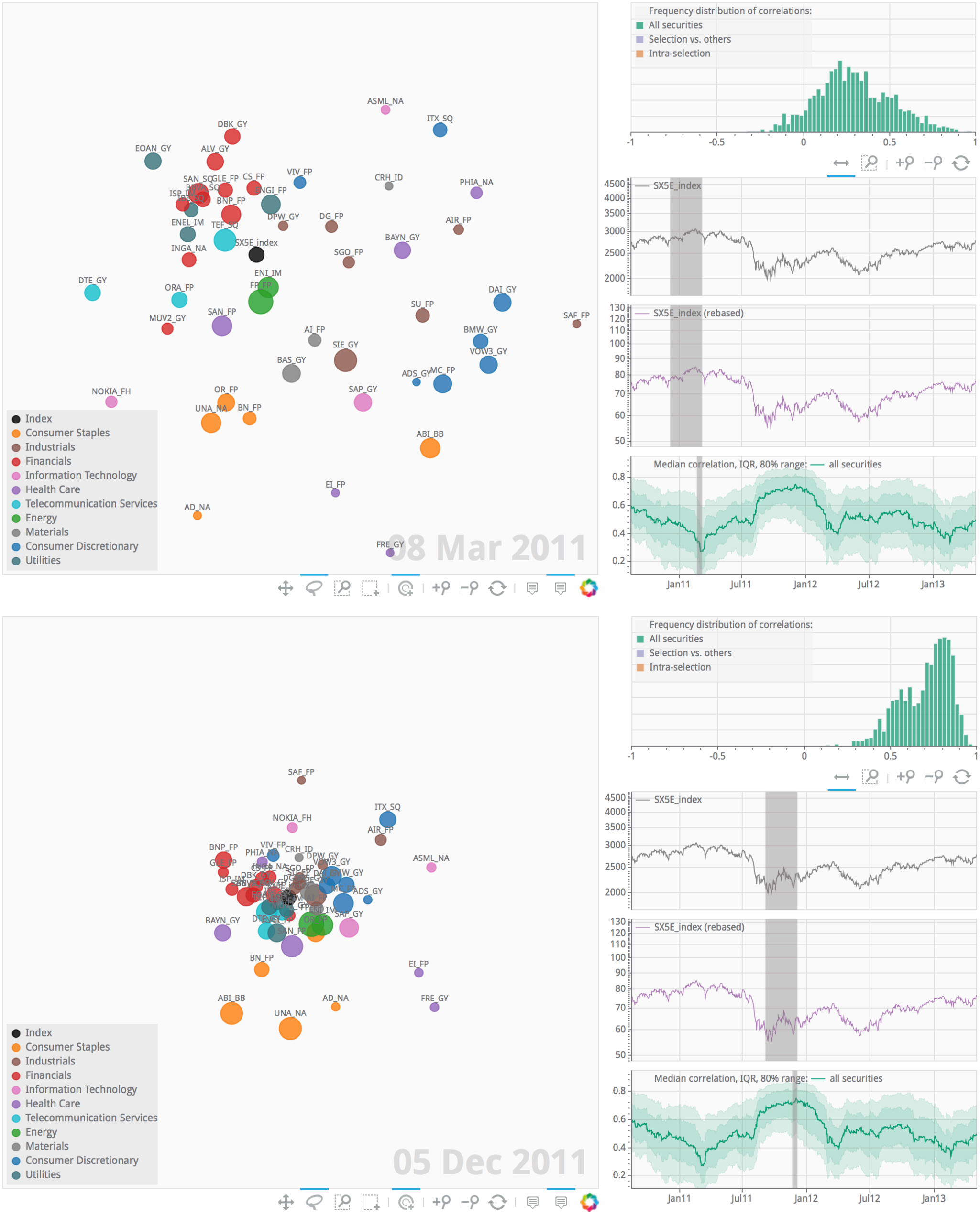


5.1.2 Overview of high level correlation behaviour and shifts in correlation structure: the Eurozone crisis of 2011

Figure 5-2

c|swarm displays for EuroSTOXX 50 constituents before and during the Eurozone debt crisis of 2011

Upper panel: Early in 2011, the index's median correlation was low; nodes are spread out across the display. Lower panel: By mid-September, correlations had increased sharply after the market's correction, and remained elevated for some time afterwards; the swarm plot shows a tightly bunched group of stocks. Selecting all nodes and filtering out edges with r<0.7 allows us to see that stocks from the Consumer Staples sector, coloured orange — often seen as a defensive 'safe haven' for investors during times of crisis — two 'secular growth' stories, Fresenius (FRE) and ASM Lithography (ASML), as well as Deutsche Telekom (DTE) were outliers, standing apart from the rest of the market.

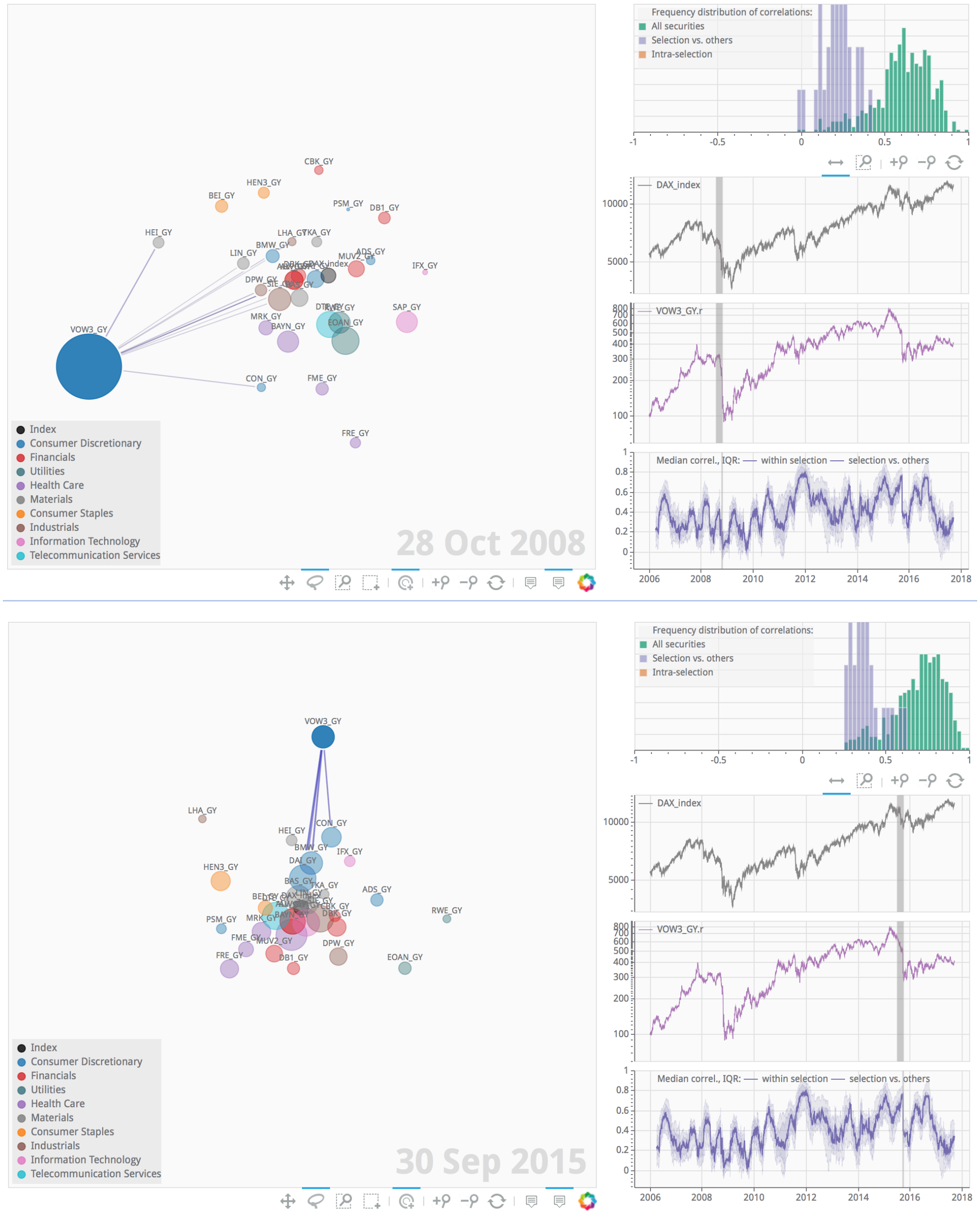


5.1.3 Identification of low-correlation stocks and outliers: Volkswagen's big moves

Figure 5-3

Volkswagen (VOW3_GY) and the DAX: two outlier situations, two very different reasons

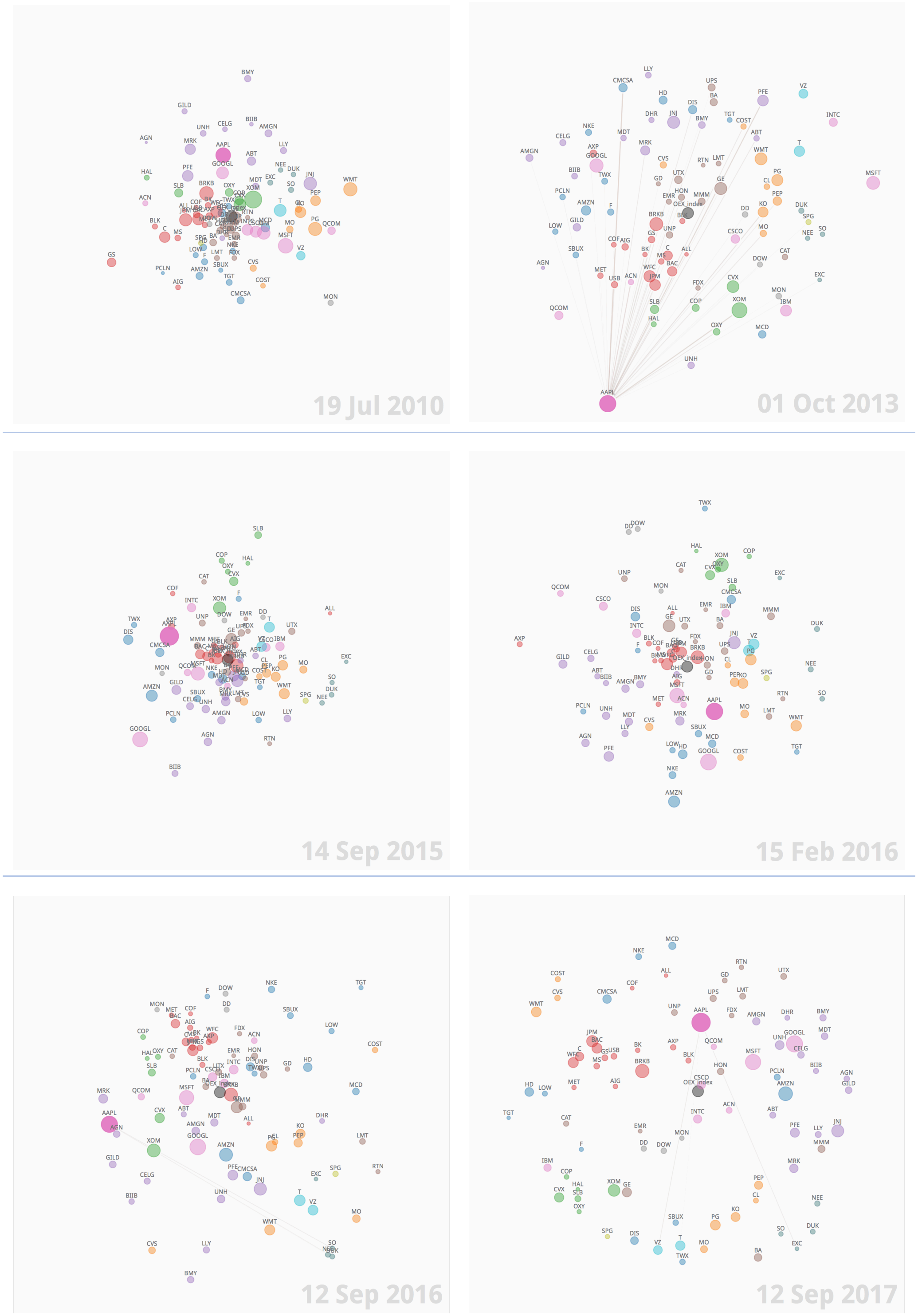
Upper panel: during 2008's global financial crisis, a short squeeze briefly made VW the world's most valuable company (hence the very large circle), sharply reducing the stock's correlations to the rest of the index at a time when other index stocks were closely correlated (VOW3 selected, edges for r<0.25 filtered out). Lower panel: in a less extreme example, in September 2015, news of the VW diesel emissions scandal broke, sending VW shares sharply lower and greatly reducing the stock's correlation to the rest of the index again (VOW3 selected, edges for r<0.5 filtered out).



5.1.4 Correlation behaviour of a single stock: Apple's shifting correlations

Figure 5-4

Apple (AAPL) correlations over time: highly variable correlation behaviour results in AAPL's node moving around a lot. *AAPL selected. Edges with r>0 filtered out. Note the changing sector colours of the stocks that AAPL is close to in the visualisation at the different points in time, and how much it moves around.*



19 Jul 2010

01 Oct 2013

14 Sep 2015

15 Feb 2016

12 Sep 2016

12 Sep 2017

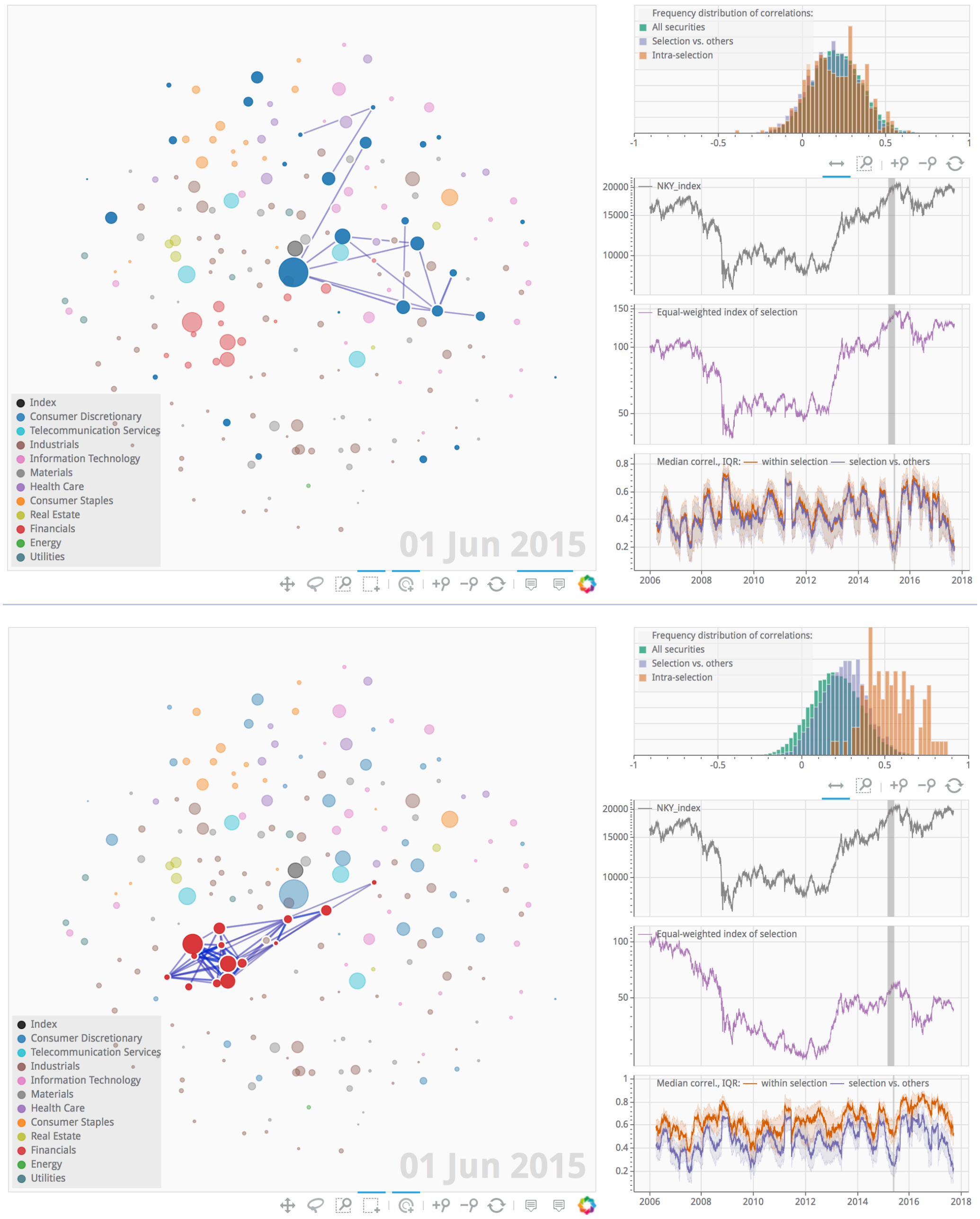
5.1.5 Visual analysis of clusters:

Japan: not all sectors are created equal, some are more equal than others

Figure 5-5

Japan: different sectors exhibit different correlation structures.

Upper panel: we can see the heterogeneous nature of the Consumer Discretionary sector (selected, links with r<0.5 filtered out) — the large group of linked stocks are all auto manufacturers or suppliers; the other dark blue stocks are from a variety of industries including media, retail and consumer electronics. Lower panel: the financial sector (same filter for links, r<0.5) is much more homogeneous, from the orange histogram and the dispersion plot in the right-hand column we can see that these stocks are consistently more highly correlated amongst themselves than to others. Note also how much worse the financial stocks have performed than the overall market — whilst the Nikkei is near its highs and the consumer discretionary stocks' index is noticeably higher than its 2006 level, the equal-weighted index of the financials stocks languishes at less than half its value from January 2006.



6 Further details on layout algorithm evaluation and design

6.1 Layout and animation calculation: evaluation method

W-1 What needs to be calculated for each time slice?

W-2 How will transitions between time slices be managed?

These design choices were determined by design choices specified in earlier work elements. Calculations for each time slice would be determined by the displays and interactions specified, with transitions between the time slices determined by which of those displays' elements were not static, whilst ensuring that the algorithm ran at a sufficient speed. As users had expressed a requirement for the software to be responsive[12], offline calculation (Beck et al., 2017) was ruled out; minimising calculation times therefore became an important consideration in development. Potential animation speed was measured during development by layout algorithm run time (in wall-clock time), i.e. 1/framerate. To achieve a 25fps frame rate (equivalent to PAL analogue television) with an acceptable overhead for other computational tasks, calculation times of 0.01 sec/iteration or less were targeted for datasets up to the maximum scalability design requirement of 500 securities.

6.1.1 Graph layout calculation Experiment I: visualisation method and algorithm—preliminary

W-3 How will the graph layout be calculated?

The optimal algorithm to calculate graph layout was determined experimentally, using Python 3.6 code drawing on the *numpy, scipy, pandas, networkx, scikit-learn* and *bokeh* libraries.

In an initial exploration, experiment I, seven different candidate algorithms (including a baseline force-directed layout and a baseline simplistic representation plotting security beta against the market index versus correlation against the market index) were selected and evaluated for calculation time, layout quality and stability of layout quality using the small IDAX development dataset. Layout quality was measured using Pearson and Spearman correlation coefficients between

- the vectorized correlation matrices (i.e. stocks' pairwise proximities), and
- arrays of pairwise Euclidean distances between the nodes in the calculated layouts,

with a high negative correlation coefficient being desirable (i.e. a strong negative relationship between the distances between pairs of nodes and the correlations of those nodes, as highly-correlated stocks should be shown as close together); this measure is conceptually similar to the *Stress-1* measure which is optimised (minimised) in multi-dimensional scaling (Borg et al., 2012), but easier to apply to other algorithms. Stability of layout quality was evaluated by plotting time-series of layout quality and calculation time per layout. Further anecdotal sense checks were carried out by visually inspecting the resulting layouts and manually checking correlations between arbitrarily selected pairs of nodes against their projected distance.

As generalisability of algorithms was not being tested here (unlike in, say, machine learning), a cross-validation or bootstrapping approach was felt to be inappropriate for the testing in this context; layout quality and its stability across a wide variety of datasets and environments was felt to be more important. To allow for an element of

[12] To inputs and interactions—i.e., *not* responsive web design

generalisability testing, previously unseen data (the INKY test dataset) was added in performance validation.

6.1.2 Graph layout calculation Experiment II: finer detail

The results from Experiment I led to six further questions for investigation:

W-4 How much do MDS performance and layout quality deteriorate for larger matrices?

W-5 Is MDS performance stable over longer time series of sample data?

W-6 Can MDS performance be improved to an extent that its use is feasible whilst still achieving a high frame rate (25fps or better)?

W-7 Can the MDS algorithm be implemented in such a way that similar correlation matrices will result in similar projections?

W-8 Can PCA or SVD performance be improved by adjusting its hyperparameters?

W-9 Can Fruchterman-Reingold performance be improved by pruning the network, using a technique akin to Threshold Networks, or by using a better implementation?

To answer these questions, a second, more detailed experiment (II) was carried out using a greater variety of longer time series, where various hyperparameters were tweaked for the more successful algorithms from experiment I. Additional statistics including mean and median absolute error and mean squared error were captured.

As the results for the Fruchterman-Reingold algorithm in experiment I were rather worse than expected, a different implementation of the algorithm was tried[13], which appeared to address some of the issues found. Performance improvement for F-R and MDS algorithms was also attempted by iterating coordinates returned by those algorithms as seeds (initial positions) for calculation of the next time slice's layouts.

6.2 Results

Several layout techniques and Python libraries were tested before a satisfactory solution was found. For each time slice, a number of arrays are calculated (assuming that the correlation cube is calculated upon opening the software):

- Coordinates of all nodes on the main display (i.e. the graph layout)
- Which links between nodes are to be displayed, given the current filter
- The node sizes (circle areas), from their market value at that point in time
- The distribution of correlation coefficients (for the histogram)
- Positions of the ancillary line plots' bands marking the current time window shown

Calculation of these arrays, other than the first, is trivial. The best algorithm for graph layout was determined experimentally.

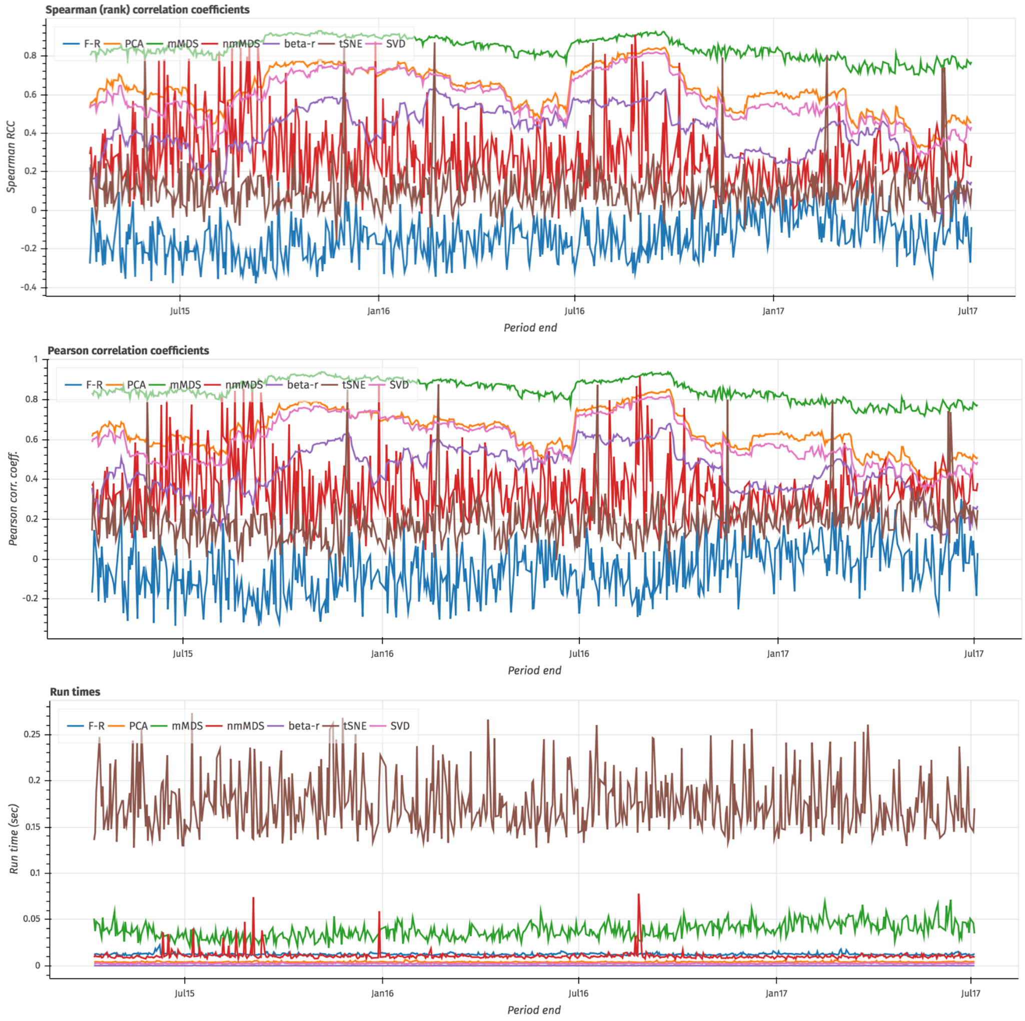
6.2.1 Experiment I: initial evaluation of candidate layout algorithms

Results are shown in Figure 6-1. Metric MDS produced the "best" layouts, performing noticeably better than any of the other layout algorithms, consistently calculating layouts with correlation coefficients of around 0.8 in the period examined, with this quality statistic being more stable over time than any other algorithm considered. At the other end of the spectrum, the Fruchterman-Reingold and t-SNE algorithms tested returned very poor layouts, with intermittent high p-values of their correlation coefficients suggesting that their performance was not statistically significantly better than randomly distributing nodes on the plots (t-SNE was also very slow). This

[13] from the pre-release build of *networkx*, v2.0rc11

impression was further reinforced by the wide dispersion of their metrics in the short period of time examined (and by visual inspection of the layouts produced).

Figure 6-1
Experiment I: layout quality and performance results for seven different layout algorithms[14]
Upper panel: Spearman (rank) correlation. Centre panel: Pearson correlation. Lower panel: run time.



Unfortunately, the MDS implementation deployed here also had two significant drawbacks which sharply reduced its utility: it was rather slow, with further tests suggesting that its performance deteriorates noticeably as correlation matrix size increases, and the stochastic nature of MDS meant that two layouts from time-adjacent similar correlation matrices typically did not look particularly alike, thus not achieving the objective of stability over time when correlation structure is not changing much.

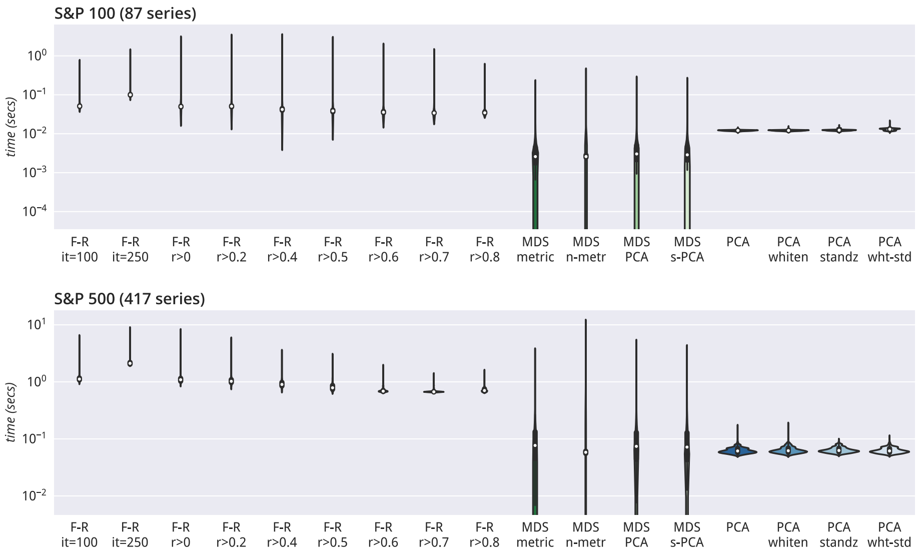
6.2.2 Experiment II: selection of layout algorithm to be used

'Seeding' the MDS layouts using the previous time slice's coordinates resulted in substantial improvements in calculation times by reducing the number of iterations needed for the algorithm to converge, usually outperforming other techniques for all data sets other than the largest (although PCA calculation times were most stable; see

[14] abbreviations: F-R: Fruchterman-Reingold; PCA: principal component analysis; mMDS: metric multi-dimensional scaling; nmMDS: non-metric multi-dimensional scaling; beta-r: simple scatter plot of stock beta vs correlation coefficient; tSNE: t-distributed stochastic neighbour embedding; tSVD: truncated singular value decomposition.

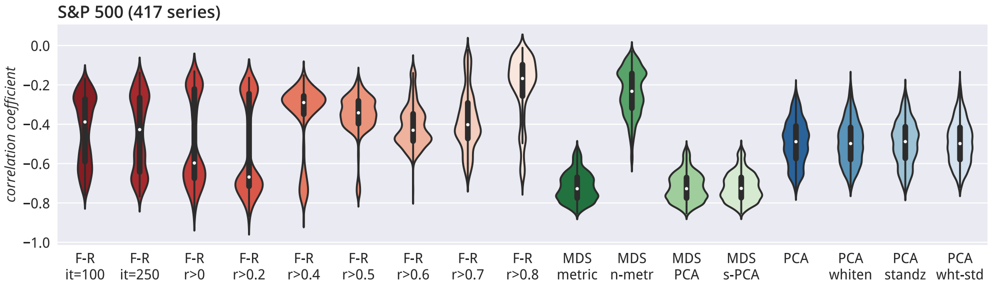
Figure 6-2). The layouts returned by MDS were of consistently higher quality than those from the other candidate algorithms, with visibly lower dispersion.

Figure 6-2
Experiment II: distribution of layout calculation times by algorithm/parameter set (selected data)



The ‘seeded’ MDS layouts also had the advantage that they were quite stable when comparing adjoining time slices, except when there was a large shift in correlation structure, in which case a noticeable change in the layout was desired. This would prove very useful for calculating frames for the animation. Layout quality appeared to be largely independent of dataset size, with the general level of correlation at a given point in time appearing to be a greater factor (for an illustration of this, see Figure 6-5). As changing other high-level MDS hyperparameters did not appear to make a noticeable difference, other than the deterioration in layout quality caused by using non-metric (ordinal) MDS, standard (metric) MDS with ‘seeding’ was selected for use as the prototype software’s layout algorithm.

Figure 6-3
Experiment II: distribution of layout quality (only the largest data set tested is shown)

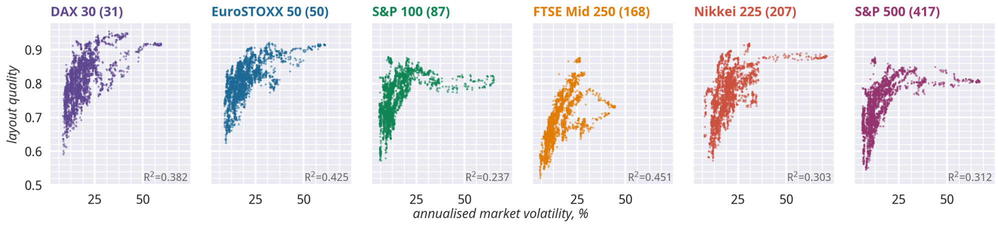


6.3 Testing, validation and evaluation

6.3.1 Validation of algorithm design and of system implementation

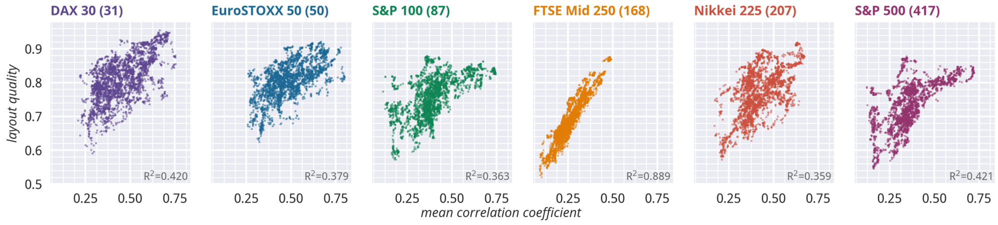
On the whole, layouts were of acceptable quality, with layout quality measures of 0.7 or better in most cases; quality tended to suffer somewhat during low-volatility periods in markets (Figure 6-4). This is consistent with the understanding that multi-dimensional scaling techniques perform better with datasets containing greater structure (Borg et al., 2012).

Figure 6-4
Layout quality vs. market volatility: by test dataset



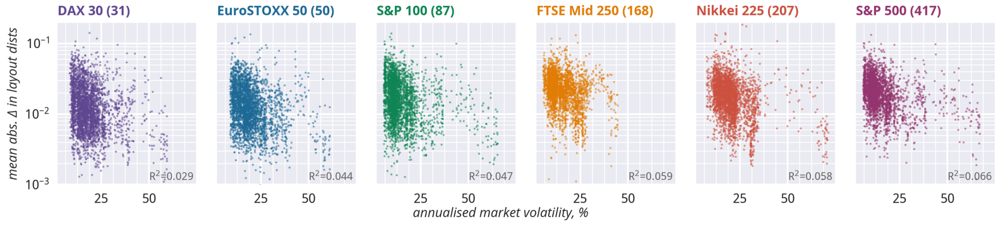
Plotting layout quality against a central tendency measure of correlation levels in the source data revealed that the relationship between layout quality and volatility could, to an extent, be explained by the presence of a confounding variable related to both, namely that correlations tended to be higher in times of high volatility (Figure 6-5).

Figure 6-5
Layout quality vs. mean correlation coefficient: by test dataset



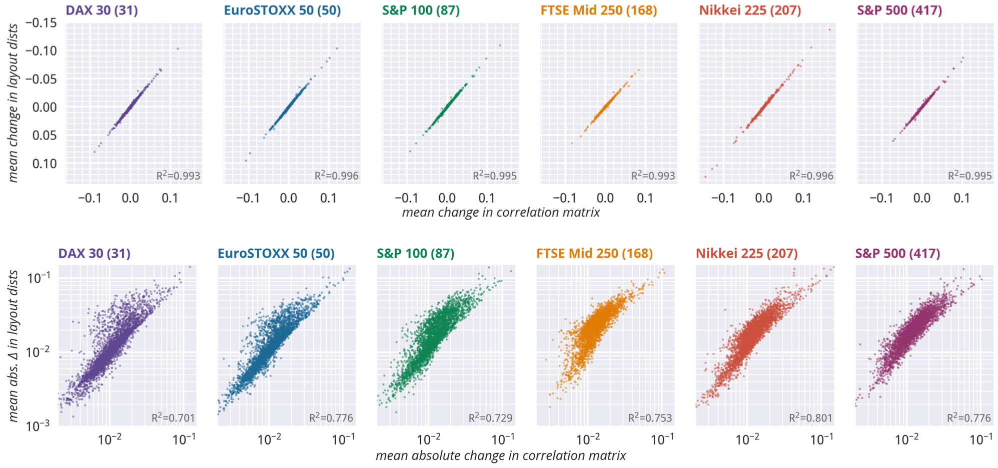
Layout stability across time was unrelated to market volatility (Figure 6-6).

Figure 6-6
Frame-by-frame layout abs. stability vs. market volatility: by test dataset



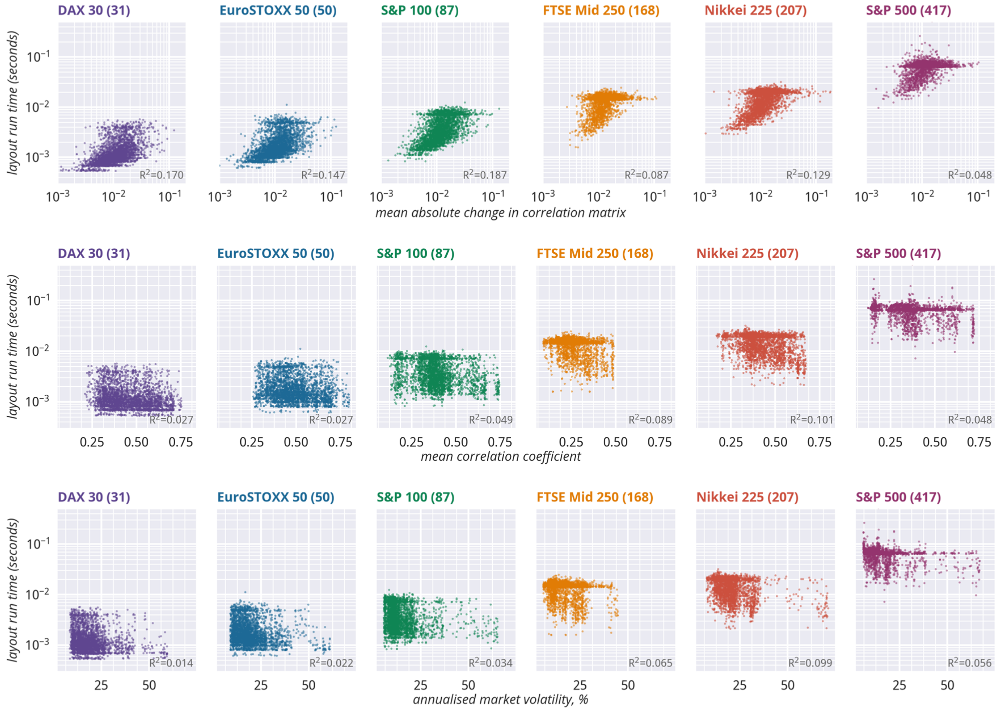
There was a strong relationship apparent between changes in the layout and changes in the ground-truth data, suggesting that times of stable correlations had successfully been encoded as drawing stability, with unstable correlations showing a significantly greater degree of movement in the course of the animation (Figure 6-7).

Figure 6-7
Upper: frame-by-frame layout stability vs. correlation matrix stability: by test dataset (*NB: inverted y-axis*)
Lower: frame-by-frame layout abs. stability vs. correlation matrix abs. stability: by test dataset



Layout calculation times (Figure 6-8) depended somewhat on the number of iterations required to achieve a stress-minimising layout, which had a weak (albeit statistically significant) relationship with the change in the ground-truth correlation matrix. Calculation times certainly varied with correlation matrix size, approximating $O(n^2/2)$ complexity (improved from $O(n^2)$ by the use of vectorization), with calculations for the S&P 500 dataset in a few instances being too slow to provide a good frame rate for animation. On the other hand, calculation times seemed to be generally independent of levels of correlations or market volatility.

Figure 6-8
Upper: layout calculation time per frame vs frame-by-frame correlation matrix stability
Middle: layout calculation time per frame vs mean correlation coefficient
Lower: layout calculation time per frame vs market volatility



6.4 Appendix: detailed Experiment II results

Figure 6-9

Performance: distribution of per-layout run time by algorithm, by data set (*lower values are better*)

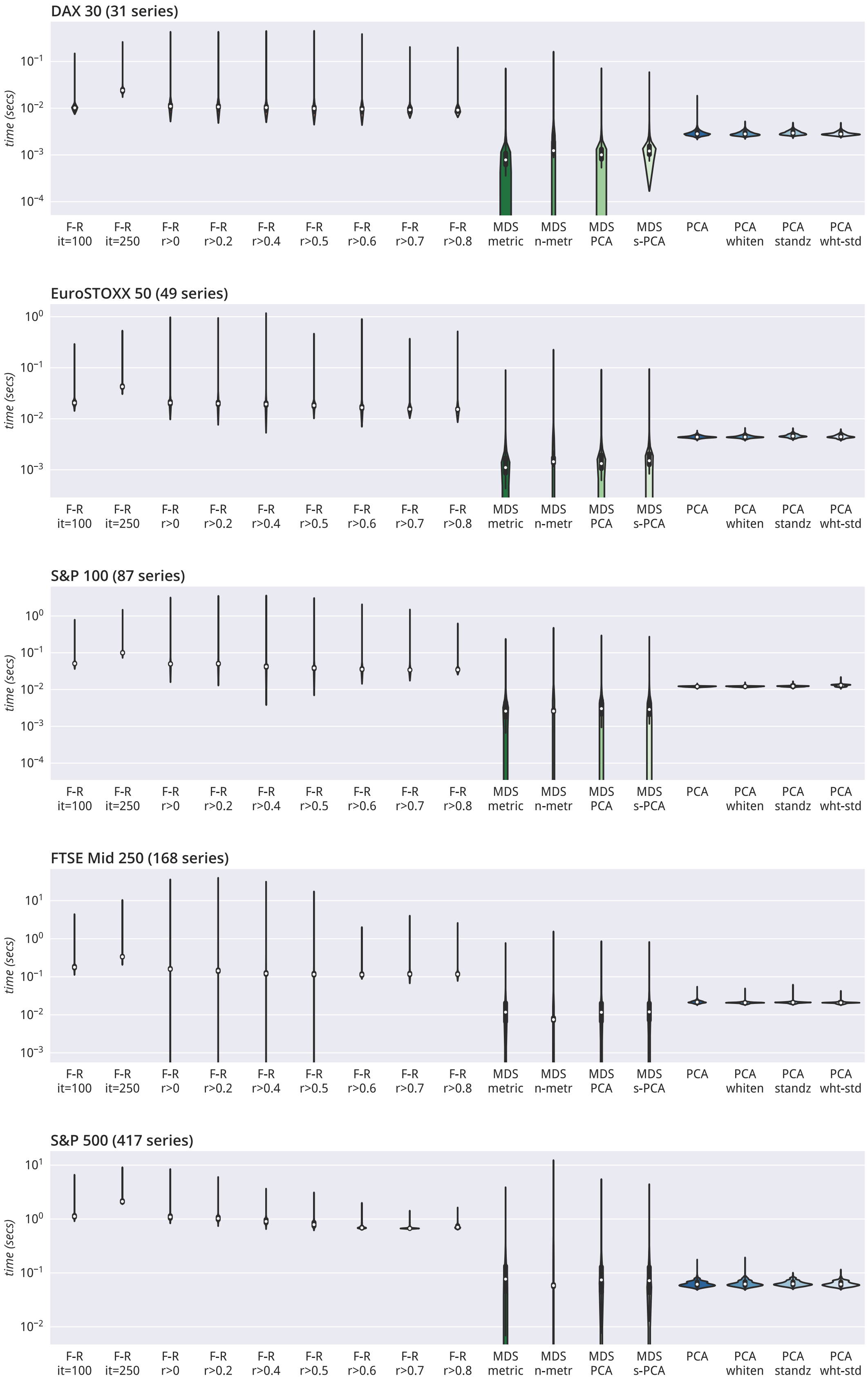


Figure 6-10

Quality: distribution of per-layout correlation coefficient by algorithm, by data set *(lower values are better)*

Correlation between correlation coefficient and pairwise Euclidean distance between layout coordinates

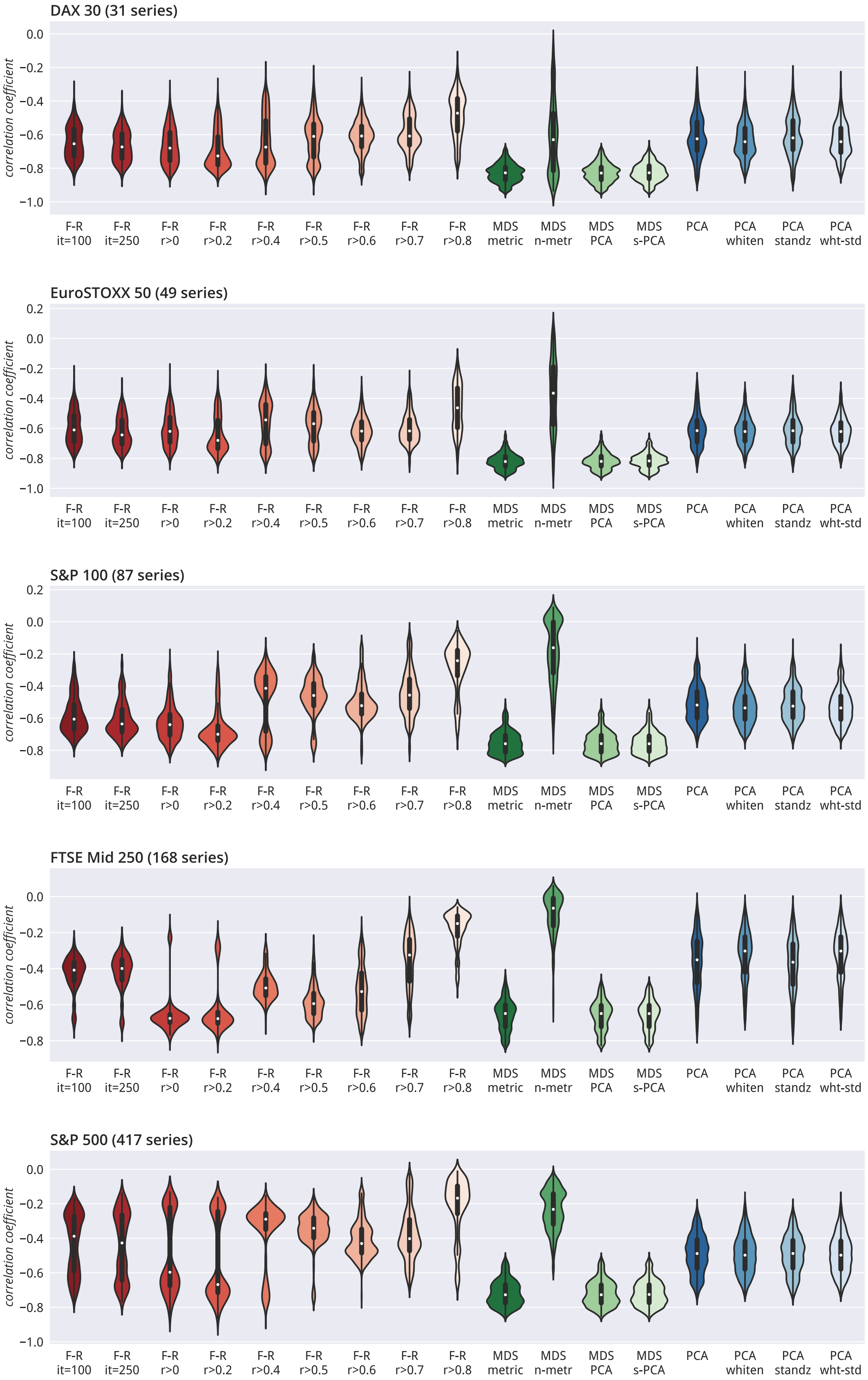


Figure 6-11

***Distribution of per-layout median absolute error by algorithm, by data set** (lower values are better)*

** layout and ground truth data rescaled to [0,1] intervals for comparability*

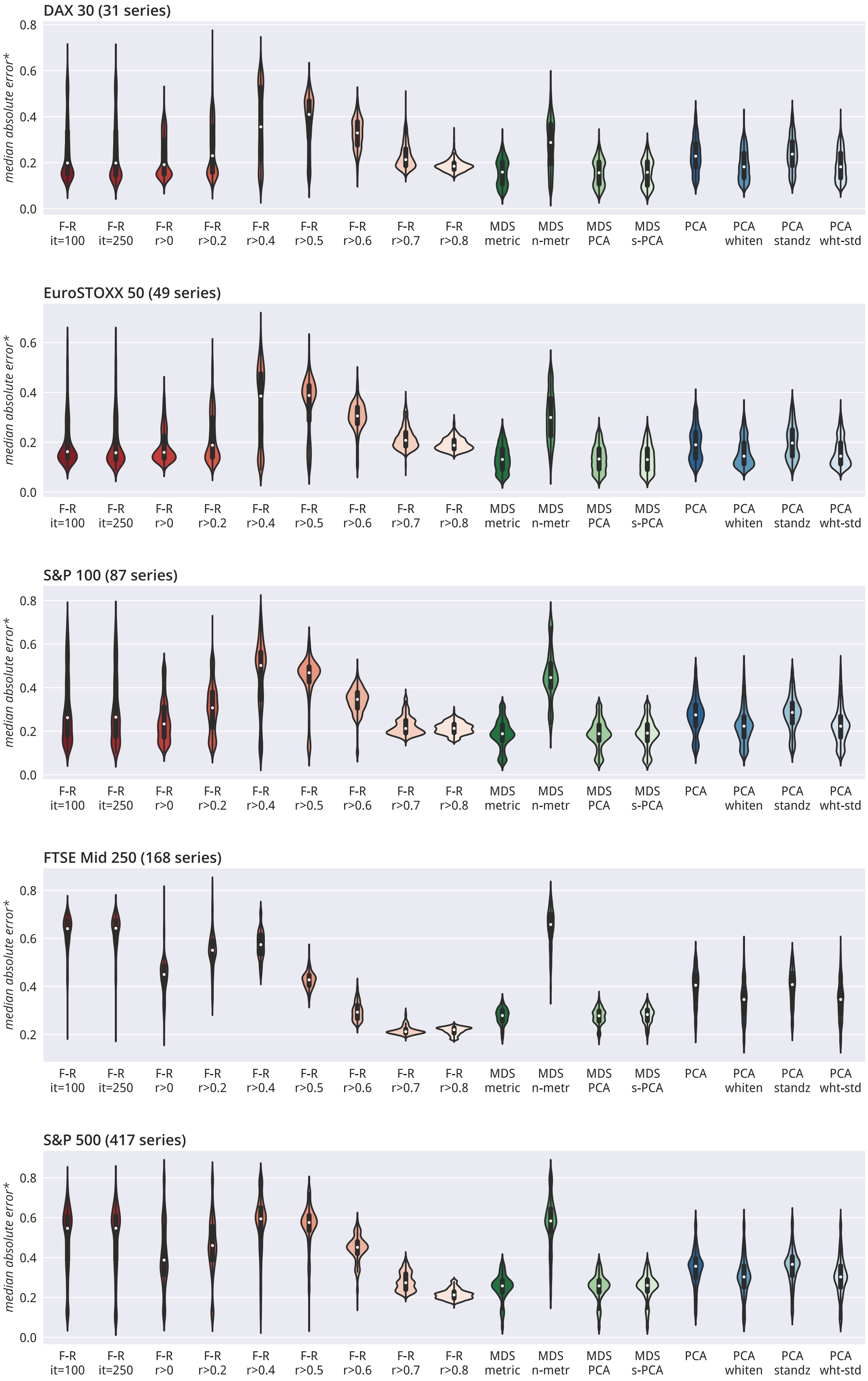
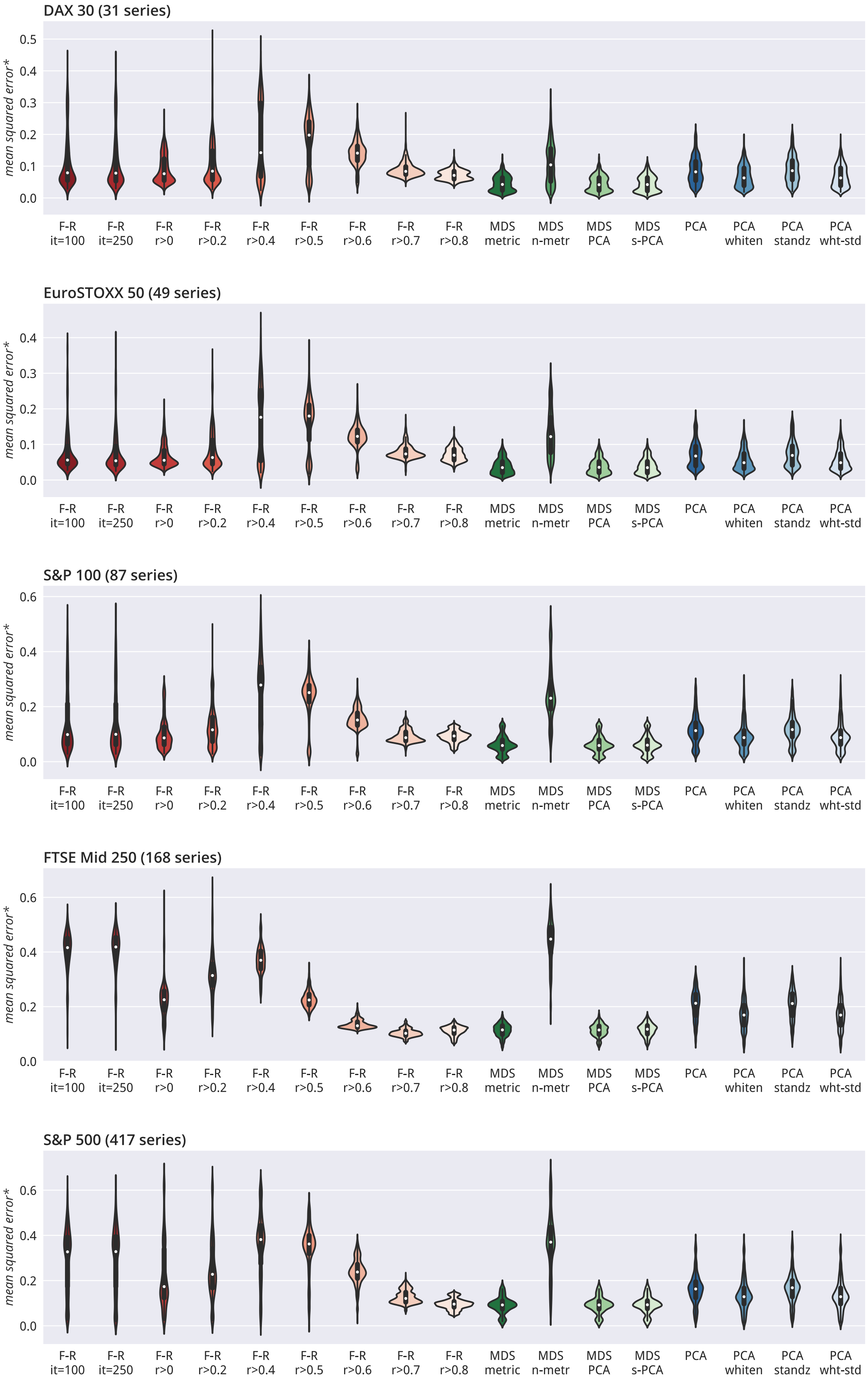


Figure 6-12

***Distribution of per-layout mean squared error by algorithm, by data set** (lower values are better)*

** layout and ground truth data rescaled to [0,1] intervals for comparability*



DAX 30 (31 series)

EuroSTOXX 50 (49 series)

S&P 100 (87 series)

FTSE Mid 250 (168 series)

S&P 500 (417 series)

7 Software copyright and license acknowledgments

c|swarm proximity swarm visualisation software:
Copyright © 2017 Scaridae Limited and Peter M. Simon. All rights reserved.
This software uses the open-source technologies listed below, the authors' intellectual property and copyrights are acknowledged.

The open-source software libraries and languages listed below are used in accordance with their licences and the intellectual property rights of their authors are acknowledged as follows:

Bokeh
Copyright © 2012-2017 Anaconda, Inc. All rights reserved.

NumPy
Copyright © 2005-2017, NumPy Developers. All rights reserved.

SciPy
Copyright © 2003-2013 SciPy Developers. All rights reserved.
Copyright © 2001, 2002 Enthought, Inc. All rights reserved.

Pandas
Copyright © 2008-2017, AQR Capital Management, LLC, Lambda Foundry, Inc. and PyData Development Team. All rights reserved.

Scikit-Learn
Copyright © 2007-2017, Scikit-Learn developers. All rights reserved.

The above packages are licensed under the 3-clause BSD license:

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

* Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

* Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

* Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Python 3.6.2
Copyright © 2001-2016 Python Software Foundation. All rights reserved.
Copyright © 2000 BeOpen.com. All rights reserved.
Copyright © 1995-2000 Corporation for National Research Initiatives. All rights reserved.
Copyright © 1991-1995 Stichting Mathematisch Centrum. All rights reserved.
Licensed under Python Software Foundation license agreement

Colorcet
Copyright © 2015 Peter Kovesi/Centre for Exploration Targeting.
Licensed under Creative Commons 4.0.

The above may incorporate copyrighted software from other parties, whose rights are acknowledged.