

Introduction to Longitudinal Data Analysis

The Workflow

Vernon Gayle


Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern
<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN




Administrative Data Research Centre Scotland

Better Knowledge Better Society

- Supporting researchers to use administrative data for economic and social research
- Enabling studies that benefit the public
- Helping inform and improve policy-making

W. adm.ac.uk
E. adrc-sped.ac.uk





AQMeN
Applied Quantitative Methods Network

Research Centre developing a dynamic engineering set of inter-disciplinary skills to improve understanding of social issues in the UK

conducting an international programme of research using secondary data in three primary areas:

- Health and Wellbeing
- Crime and Social Stratification
- Segregation and Inequality

engaging policy makers and partners with robust independent evidence

providing high quality statistical analysis and user support

facilitating knowledge exchange between academic disciplines across sectors

improving quality in quantitative research in the UK



Administrative Data Research Centre Scotland

An ESRC Data Investment

Better Knowledge Better Society

- Supporting researchers to use administrative data for economic and social research
- Enabling studies that benefit the public
- Helping inform and improve policy-making



W. adm.ac.uk
E. adrc-sped.ac.uk

In partnership with the First Institute of Scotland and the Scottish Government



AQMeN
Applied Quantitative Methods Network



A Research Centre developing a pioneering set of inter-disciplinary skills to improve understanding of social issues in the UK

conducting an international programme of research using secondary data in three primary areas:

- Health and Wellbeing
- Crime and Social Stratification
- Segregation and Inequality

engaging policy makers and partners with robust independent evidence

providing high quality statistical analysis and user support

facilitating knowledge exchange between academic disciplines across sectors

improving quality in quantitative research in the UK



A Thought Experiment

Be honest....

1. Have you ever lost a file?
2. Have you ever wondered if you have deleted a file?
3. Have you and a colleague ever been working on different versions of a file?

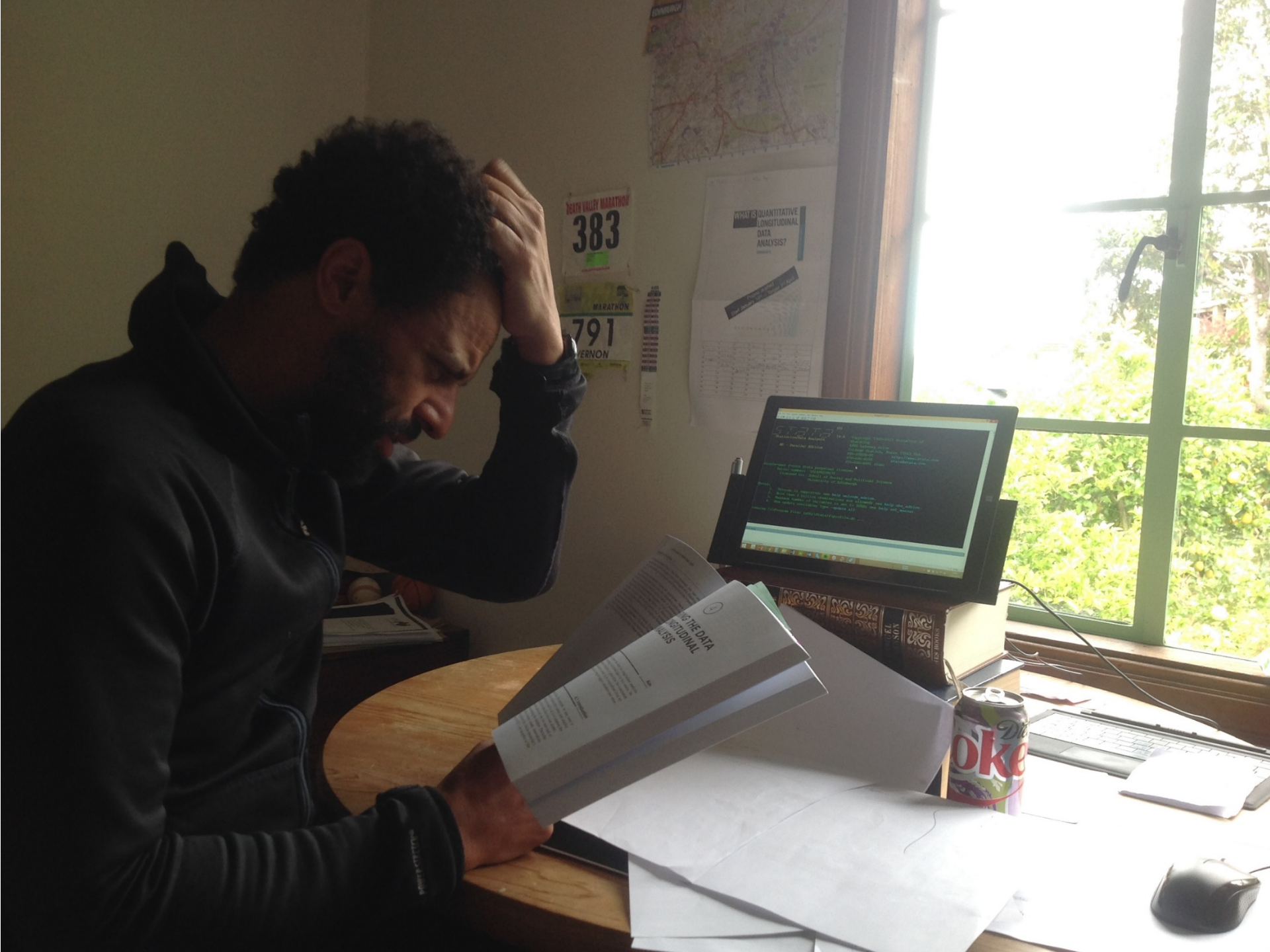
A Thought Experiment

Be honest....

Have you ever struggled to identify which data file is the correct one?

chapter1_2014.dat

chap1_2014.dat





The Workflow

- Planning, organising and documenting work
- This includes...
 - Cleaning data*
 - Analysing data*
 - Presenting results*
 - Backing up and archiving material*

The Workflow

Workflow should be planned and carefully orchestrated

Workflow MUST not be *adhoc*
(e.g. piece-meal, developed as a reaction to mistakes etc.)

The Workflow

Better supporting YOU and what YOU DO

Not changing you into something YOU ARE NOT

Shopping without a list?
Cooking with a list?



APPROVED B-17F and G CHECKLIST

REVISED 3-1-44

PILOT'S DUTIES IN RED

COPILOT'S DUTIES IN BLACK

BEFORE STARTING

1. Pilot's Preflight—**COMPLETE**
2. Form 1A—**CHECKED**
3. Controls and Seats—**CHECKED**
4. Fuel Transfer Valves & Switch—**OFF**
5. Intercoolers—Cold
6. Gyros—**UNCAGED**
7. Fuel Shut-off Switches—**OPEN**
8. Gear Switch—**NEUTRAL**
9. Cowl Flaps—Open Right—**OPEN LEFT**—Locked
10. Turbos—**OFF**
11. Idle cut-off—**CHECKED**
12. Throttles—**CLOSED**
13. High RPM—**CHECKED**
14. Autopilot—**OFF**
15. De-icers and Anti-icers, Wing and Prop—**OFF**
16. Cabin Heat—**OFF**
17. Generators—**OFF**

STARTING ENGINES

1. Fire Guard and Call Clear—**LEFT** Right
2. Master Switch—**ON**
3. Battery switches and inverters—**ON & CHECKED**
4. Parking Brakes—Hydraulic Check—**On-CHECKED**
5. Booster Pumps—Pressure—**ON & CHECKED**
6. Carburetor Filters—Open
7. Fuel Quantity—Gallons per tank
8. Start Engines: both magnetos on after one revolution
9. Flight Indicator & Vacuum Pressures—**CHECKED**
10. Radio—On
11. Check Instruments—**CHECKED**
12. Crew Report
13. Radio Call & Altimeter—**SET**

ENGINE RUN-UP

1. Brakes—Locked
2. Trim Tabs—**SET**
3. Exercise Turbos and Props
4. Check Generators—**CHECKED & OFF**
5. Run up Engines

BEFORE TAKEOFF

1. Tailwheel—Locked
2. Gyro—Set
3. Generators—**ON**

AFTER TAKEOFF

1. Wheel—**PILOT'S SIGNAL**
2. Power Reduction
3. Cowl Flaps
4. Wheel Check—OK right—**OK LEFT**

BEFORE LANDING

1. Radio Call, Altimeter—**SET**
2. Crew Positions—OK
3. Autopilot—**OFF**
4. Booster Pumps—On
5. Mixture Controls—**AUTO-RICH**
6. Intercooler—Set
7. Carburetor Filters—Open
8. Wing De-icers—Off
9. Landing Gear
 - a. Visual—Down Right—**DOWN LEFT**
Tailwheel Down, Antenna in, Ball Turret Checked
 - b. Light—**OK**
 - c. Switch Off—Neutral
10. Hydraulic Pressure—**OK** Valve closed
11. RPM 2100—Set
12. Turbos—Set
13. Flaps $\frac{1}{2}$ — $\frac{1}{2}$ Down

FINAL APPROACH

14. Flaps—**PILOT'S SIGNAL**
15. RPM 2200—**PILOT'S SIGNAL**

In the late 1930s, military aviators in the American Army and Navy began using aviation checklists. Checklist became part of a new paradigm for how to fly, which consisted of

- Elaborate standardized procedures for many activities
- Checklists to ensure all critical steps had been done
- Quantitative tables and formulas that specified the best settings, under different conditions, for speed, engine RPM, gasoline/air mixture, engine cooling, and many other parameters.

This new paradigm (Standard Procedure Flying) had a major influence on reducing aviation accidents and increasing military effectiveness during World War II, particularly because of the rapidly increasing complexity of military aircraft, and the huge number of new pilots.

Despite the benefits of Standard Procedure Flying for both safety and efficiency, by the end of WWII only a few air forces had fully embraced it

Roger Bohn <http://www.vs29.org/Links/NATOPS/SOP-bohn-2013-1.pdf>

The Workflow

The Platinum Standard



The Four Pillars of Wisdom



A Planned Workflow Has Benefits



A stylized illustration of a classical building facade with four columns. Each column is inscribed with a benefit of a planned workflow. The columns are white with vertical fluting and are set against a dark gray background. The building has a triangular pediment and a base.

Accuracy

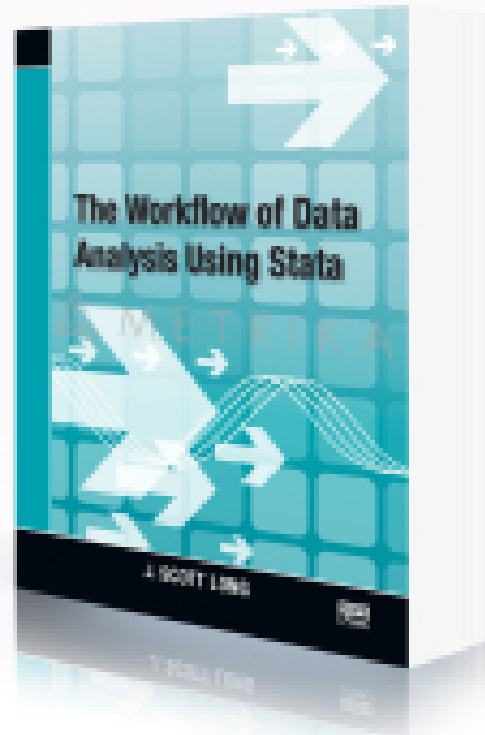
Progamming
Efficiency

Transparency

Reproducibility

Four Pillars of Wisdom

- Accuracy
 - minimising information loss and errors in analyses and output
- Programming Efficiency
 - automation, maximising features in software
- Transparency
 - showing what you did, why, when, how
- Reproducibility
 - same results every time whoever or wherever
 - editing, rewriting reports or re-submission of papers



The best habit that you can get into



is to get into good habits!

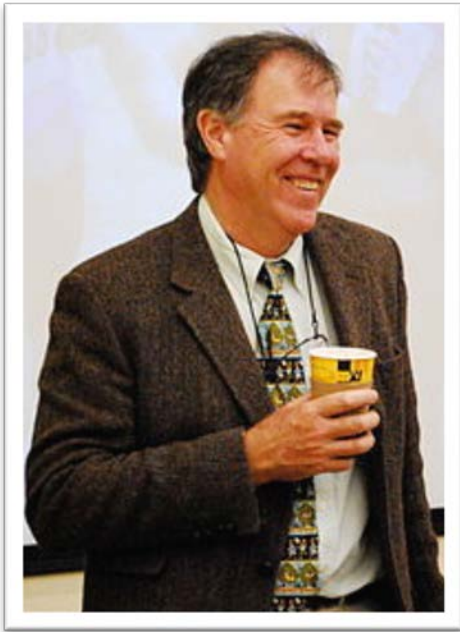
The Workflow

Drukker's dictum: Never type anything that you can obtain from a saved result

My dictum (Gayle's dictum):

You can't be too fit or have too many publications

However...



- 500+ scientific publications in peer reviewed journals (15,000+ citations and an H-index of 66)
- Has run more than 70 marathon and ultra-marathon races, including seven 90km Comrades Marathons and fifteen 56km Two Oceans Marathons

http://www.essm.uct.ac.za/ESSM/Tim_Noakes



- Over 20 Ultra Marathons including the Western States 100 mile race
- 1480 citations since 2011

<https://www.stat.berkeley.edu/~stark/index.html>

Long's Law

It is always easier to document today than it is tomorrow!

Corollary 1:

Nobody likes to write documentation

Corollary 2:

Nobody every regrets having written documentation

Long's Law

Has anyone in the history of data analysis ever said

“these files are too well documented”





As with many scientists, Linus Pauling utilized bound notebooks to keep track of the details of his research as it unfolded. A testament to the remarkable length and diversity of Dr. Pauling's career, the Pauling Papers holdings include forty-six research notebooks spanning the years of 1922 to 1994 and covering any number of the scientific fields in which Dr. Pauling involved himself. In this regard, the notebooks contain many of Pauling's laboratory calculations and experimental data, as well as scientific conclusions, ideas for further research and numerous autobiographical musings.

Research Notebook 01

1922

Research Notebook 02

1922-1923, 1932, 1934, 1936, 1973,
1985

Research Notebook 03

1923-1925

Research Notebook 04

1923-1924, 1928-1930

Research Notebook 05

Research Notebook 13

1935-1936, 1938-1939

Research Notebook 14

1936-1939, 1949, 1952

Research Notebook 15

1935, 1937, 1968

Research Notebook 16

1935-1956

Research Notebook 17

1939-1941, 1971, 1988

Research Notebook 24

1953, 1956, 1962, 1963, 1967, 1968,
1969, 1970, 1973

Research Notebook 25

1958, 1964-1966

Research Notebook 26

1955, 1964-1969, 1974-1976, 1980-
1982, 1987, 1990-1991

Research Notebook 27

1952-1954, 1960-1961, 1964, 1971-

Research Notebook 35b

1938-1939, 1946, 1955, 1968, 1986-
1988

Research Notebook 36

1980-1981, 1986-1987

Research Notebook 37

1971, 1983

Research Notebook 38

1980-1981, 1983, 1985, 1989

Research Notebook 39

[Previous: 150](#)Book: 24 ▼ Page: Go

24 June 1973
Portola Valley, Cal. Golden Wedding Anniversary 150
Luis Pauling
Three days ago Ava Helen and I
celebrated our 50th wedding anniversary. We
had been married in Salem, Oregon, on
17 June 1923.

Our celebration was at the ranch. Our
four children were there, also Linda's
husband (Barclay) and their four boys;
also Art and Lawrence Robinson; also Art
Cherkin; also Ewan Cameron from Scotland,
his wife, and two children; also
L. Jorge Miller, Ava Helen's sister;
also my two sisters, Pauline Pauling-Ney
and Lucille Jenkins

The Workflow

- Improving the workflow with a modest amount of effort
- The less experience you have the better
 - start from the very beginning

The Workflow

ALL SERIOUS WORK MUST BE REPRODUCIBLE!

There MUST be an audit trail

The Workflow

Why is it all so difficult?

Social science data tends to come in messy formats

Administrative data often is even more complex in nature than social survey data

The Workflow

Why is it all so difficult?

Minor decisions have major consequences...

Which cases?

Which variables?

How to code (e.g. education)?

How to recode?

Where do I truncate?

The Workflow

Minor decisions have major consequences...

Which cases?

Which variables?

How to code (e.g. education)?

How to recode?

Where do I truncate?

Can I trace these decisions in my audit trail?


```
1  STOP
2
3  /**
4
5  ****
6
7  Next Actions:
8
9
10
11
12  Author:
13
14
15  Project:
16
17
18  Sub-project:
19
20
21  Date of Next Meeting (or supervision):
22
23
24  Latest Update:
25
26
27  Previous Updates:
28
29
30
31  Useful information:
32  http://www.samaritans.org/ (08457 90 90 90)
33
34
35  ****
```

A clear and consistently well organised and annotated .do file is central to successful quantitative longitudinal data analysis

It is possible to save a file called 'template.do'

A clear and consistently well organised and annotated .do file is central to successful quantitative longitudinal data analysis.

It is possible to save a file called 'template.do'

into your home Stata folder so that a blank .do file that is pre-populated with organisational information is automatically generated when you open Stata

This is easily achieved by adding a line to your Stata profile (profile.do) which points to the template

(e.g. doedit "C:\Program Files (x86)\Stata14\template.do")

The Workflow

File Naming Protocols

File Name =

name_date_depositor's initials_version_type

The Workflow

File Naming Protocols

File Name = name_date_depositor's initials_version_type

Therefore **bhpsaindresp_20140506_vg_v1.dta**

Would be a

- a.. The British Household Panel Survey File “aindresp”
- b.. Deposited on 6th May 2014
- c.. Deposited by vg (Vernon Gayle)
- d.. Version v1
- e.. File type (e.g. a Stata .dta file)

	A	C	D	E	F	G	H	I
1	File Register							
2								
3								
4		File Name (name_subname_date[year/month/day]_depositor's initials_version_type)	File Type					Brief Description of the file and its purpose
5	Directory Name	(e.g. bhps_aindresp_140129_vg_v1.dta)	(e.g. Stata data file)	Name of Author	Initials of Author	Date of Creation	Date of last revision	(e.g. Stata .do file MSc dissertation; Draft Chapter 1 PhD)
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								

Other seemingly small issues such as ‘Directory Structures’ and ‘Variable Naming Conventions’ are similarly worth thinking about!

The Workflow

Why is it all so difficult?

Poor discipline and insufficient documentation

Estimating Work Time...





Professor Vernon Gayle

vernongayle

I am Professor of Sociology and Social Statistics at the University of Edinburgh.

Edit bio

👤 University of Edinburgh
📍 Edinburgh, Scotland, UK
✉ vernongayle@ed.ac.uk
🌐 <http://www.vernongayle.com/>

Overview

Repositories 18

Stars 2

Followers 5

Following 0

Pinned repositories

Customize your pinned repositories

≡ [vernongayle.github.io](#)

Github Pages (a summary and profile)

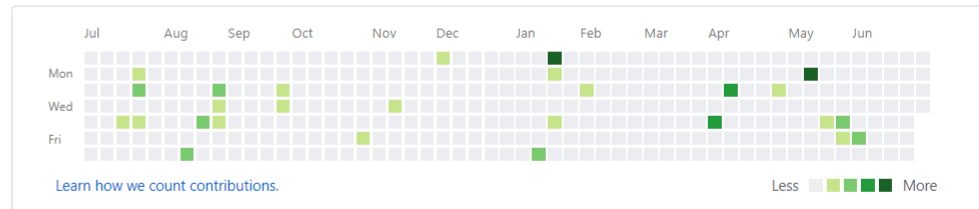
≡ [spring_into_longitudinal_data_analysis](#)

Spring into Longitudinal Data Analysis

★ 1

106 contributions in the last year

Contribution settings ▾



Contribution activity

Jump to ▾

July 2018

vernongayle has no activity yet for this period.

2018

2017

2016



GOOD LUCK!

Aim for Gold in your work!

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern
<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN