gravitas: exploring probability distributions for bivariate temporal granularities

## **Student information**

Name: Sayani Gupta

University: Monash University, Australia

Email: [gupta.sayani@gmail.com](mailto:gupta.sayani@gmail.com)

Student Email : [Sayani.gupta@monash.edu](mailto:Sayani.gupta@monash.edu)

Github: <https://github.com/Sayani07>

Twitter: <https://twitter.com/SayaniGupta07>

Timezone: AEST (UTC + 11:00)

## Abstract

Project gravitas aims to provide methods to operate on time in an automated way, to deconstruct it in many different ways. Deconstructions of time that respect the linear progression of time like days, weeks and months are defined as linear time granularities and those that accommodate for periodicities in time like hour of the day or day of the month are defined as circular granularities or calendar categorizations. Often visualizing data across these circular granularities are a way to go when we want to explore periodicities, pattern or anomalies of the data. Also, because of the large volume of data in recent days, using probability distributions for display is a potentially useful approach.The project will provide these techniques into the tidy workflow, so that probability distributions can be examined in the range of graphics available in the ggplot2 package.

## Motivation

Considerable data is accumulated by sensors today. An example is data measuring energy usage on a fine scale using smart meters. Smart meters are installed on many households in many countries now. Providing tools to explore this type of data is an important activity. Probability distributions are induced by various aggregations of the data, by temporal components, by spatial region or by type of household. Visualizing the probability distribution of different households across different circular granularities will help energy retailers find similar households or understand consumer behaviour better and consequently increase efficiency in planning ahead.

## Project gravitas

Project gravitas will consist of three parts:

1. **R Package**: Develop an R package consisting of the following three modules:

The module **BUILD** will provide the methods to exhaustively construct any granularities. For example, if the data is available at a half hourly scale, this module would create functions to compute time indices that can be constructed in combination with half-hour like half-hour of the day or half-hour of the week. Additionally, it should also be able to provide functions to create granulartities in combination with temporal scales that are one or multiple levels above half-hour like hour, day, week or month. The idea here is to able to provide exhaustive set of granularities to figure out periodicities, patterns or anomalies across different granularities that are not obvious and hence aid the process of looking at data from multiple perspectives.

The module **COMPATIBLE** will provide automatic checks on the feasibility of plotting or drawing inference from two granularities together. The idea is to categorize pairs of granularities as either a harmony or clashes, where harmonies are pairs of circular granularities that aid exploratory data analysis. Clashes are pairs that are incompatible with each other for exploratory analysis.

The module **POINTERS** will provide appropriate data structures to visualize with the grammar of graphics for harmonies. Moreover, it will provide suggestions on the nature of harmonies, which may include number of observations per combination of categories for these harmonies or variation in number of observations across combinations of harmonies. These suggestions will serve as a guide to users who are looking to explore distributions of variables across these harmonies.

1. **Shiny UI**: Develop an user interface using RShiny to enable user to walk through different modules of the package
2. **Application**: Provide examples of probability visualization of smart meter data collected on Australian households
3. **Vignette**: Document the R package functionality in a vignette

## Related work

* *lubridate* is an R package that makes it easier to work with time and also has functions for creating calendar categorizations like hour of the day, day of the week, minutes of the hour. But it mostly creates calendar categorizations that are one step up. The proposed R package will allow creating calendar categorizations that are more than one step ahead, for example, hour of the week or one step up that are not present in lubridate package like week of the month.
* Calendar based graphics in the package *sugrrants* help explore data across linear time granularities in a calendar format, whereas this package would help explore circular time granularities.
* *ggplot2* facilitates the process of mapping different variables to a 2D frame through grammar of graphics. But it does not tell us which all variables to plot together to promote exploration of data. The proposed package would provide the list of harmonies given a time variable.
* This will use as inputs *tsibble* objects which complement the tibble and extend the tidyverse concept to temporal data.

## Brief Timeline

* (Phase 0) 11 Apr - 6 May: Pre-GSoC Period
* (Phase 1) 7 May - 26 May: Community Bonding Period
* (Phase 2) 27 May - 23 June: Coding Period 1
* (Phase 3) 24 June - 28 June: Phase 1 Evaluations
* (Phase 4) 29 June -22 July: Coding Period 2
* (Phase 5) 23 July - 26 July: Phase 2 Evaluations
* (Phase 6) 27 July - 18 August: Coding Period 3
* (Phase 7) 19 August - 26 August: Final week of submitting finished product
* (Phase 8) 26 August - 2 September: Final evaluation

## Detailed Project Timeline

### Phase 0: Pre-GSoC Period

#### 3 Weeks (11 Apr - 6 May)

### Phase 1: Community Bonding Period

#### 2 Weeks (7 May - 26 May)

During the community bonding, the main focus will be to frame a road map for the project with the guidance of the mentor and improving bonding with the community.

Each module of the package would require the output from the previous one. Hence, it is important to brainstorm and map out the set of functions to be developed for each module. This will facilitate the integration of various parts of the project into one another. I will ensure all the ideas developed at this stage are implemented by the end of the GSoC period. The proposal needs to be revised at this stage to add more details on the planned activities of each of the modules.

### Phase 2 : Coding Period 1

#### 3 Weeks (27 May - 16 June)

Build functions for module **BUILD** and **COMPATIBLE**. Conduct unit tests to check robustness of functions and include documentation for these two modules. Include real world example data sets to demonstrate the utility of each of the functions functions in each module.

#### 1 Week (17 June - 23 June)

This period will be used as a buffer to complete documentation, fix bugs in the program and making the code more efficient.

### Phase 3 : GSoC Phase 1 Evaluations

This period will be used to write a detailed report on the work done in Coding Period 1. Final report will be uploaded at the project’s wiki page after feedback from mentors.

### Deliverables

* Write functions for modules **BUILD** and **COMPATIBLE**
* Full documentation regarding usage and code.
* Tests for modules **BUILD** and **COMPATIBLE**

### Phase 4: Coding Period 2

#### 2 Weeks (29 June - 13 July)

Build functions for module **POINTERS**. Conduct unit test to check robustness of functions and include documentation. Include real world example data sets to demonstrate utility of functions developed in this module.

#### 2 Weeks (14 July - 26 July)

Develop shiny UI to guide an user to navigate through the three modules of the R package.

### Phase 5: Phase 2 Evaluations

This period will be used to write a detailed report on the work done in Coding Period 2 and also a buffer to create user documentation for the shiny app. Final report will be uploaded at the project’s wiki page after feedback from mentors.

### Deliverables

* Write functions for module **POINTER** and construct unit tests
* Full documentation regarding usage and code
* shiny UI for creating a menu driven front end of gravitas

### Phase 6: Coding Period 3

#### ~ 2 Weeks (27 July - 6 August)

Document the R package functionality in a vignette

#### ~ 1 Weeks (7 August - 13 August)

Provide examples of probability visualization of smart meter data collected on Australian households

#### ~ 1 Week (14 August - 18 August)

Work on improving the shiny UI that is developed in phase 2 coding. Also document ideas that were discussed after community bonding period and could not be implemented due to time limitations. These can then form the basis of future extension of this work.

### Phase 7: Final Evaluation

All comments from mentors to be included in the final report stitching all the independent parts together and to be uploaded.

### Deliverables

* R package vignette
* Application on smart meter data
* Final report detailing all work done in GSoC period.

## Additional information regarding timeline

* The above timeline is tentative and gives a rough idea of my planned project work. I’ve no major commitments during summer (winter in Australia) and hence, will be able to dedicate 40 hours to 50 hours a week. During the last month of the project, my university will reopen and I’ll be able to dedicate around 30 hours a week. I plan to complete major chunk of the work before university reopens.
* Each week, time will be divided amongst planning, learning and coding, testing features and documentation. I plan to work on the documentation at the end of each week to minimize bugs in later stages and also establish the utility for all functions (developed that week) using small data sets from real world application.
* I’ll publish blogs at the end of each coding phase that will include highlights of the development process, hurdles that I came across and how I overcame them. Also, I would document some good practices that I learnt while looking at codes from developers of R community who are working in similar fields.
* I’ll be setting up weekly meetings with my mentors where I update them on where I am on the project with the documentation that I plan to work on at the end of each week.

## Mentors

* Dianne Cook [dicook@monash.edu](mailto:dicook@monash.edu)
* Antony Unwin