Number, Clump and Spread

P. Penguin

December 31, 2024

Abstract

We presently look at ways of numerically analysing distributions within sequences. Beginning with the problem of what feels like a failure of the timetabling at my school, this work walks through the process of developing a statistical method of sequence analysis. Summing the squares of the distances between occurrences of a limited set of recurring elements gives a measure of "spread" and summing the reciprocal of these squares gives a measure of "clump". We presently give a small number of examples of how these metrics could be implemented and interpreted.

Contents

1 Introduction

When I received my timetable at the end of the summer holidays, I was blown away by the distribution of my German lessons. With one teacher, F (in purple in figure 1), we had all our lessons in four consecutive days out of the two-week rota. This included, to the dismay of the class and the teacher, a triple period that filled up the entire space between morning break and lunch.

I allocated my free periods into blocks (in a light grey) and then created a small table of lessons and free period blocks, labelling each with "number", "clump" and "spread". This was for the purpose of planning my time: the vision was that I'd look at the "number", "clump" and "spread" of my four blocks of frees and decide to do one thing (e.g. Physics revision) in one block and another (e.g. writing hand-wavey musings in Overleaf) in another block. Remember, this was at the beginning of the year when I could be bothered with such silliness as ✨time planning✨.

But this did give me something to think about as I sat bored in physics: is there a way to formalise the thoughts I had at the beginning of the year? This entire document is essentially an exploration of my various different hunches, some of which turned out to reveal some interesting results and some of which turned out to be completely made up.

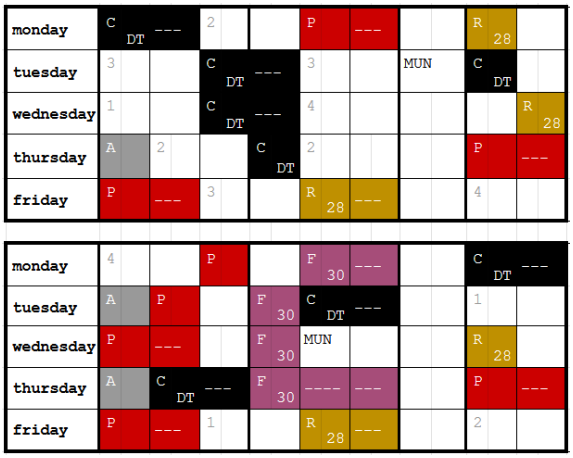

Figure 1: My 2024-25 timetable, colour-coded

2 Some Setup

We begin presently with some definitions:

$$\text{A sequence } L = L_1, L_2, \dots \text{ where } L_i \in \mathbb{L} \text{ for finite } \mathbb{L}.$$

Then we define the "sequence of distances" of L:

$$\Delta(L, l) : l \in \mathbb{L} \text{ is such that } \Delta(L, l)_i = j - k \tag{1}$$

$$\text{where } L_j = L_k (= l) \text{ are the } i\text{th and } (i-1)\text{th occurrences of } l \in L$$

This gives that the number of differences is the same as the number of occurrences of l in L:

$$|\Delta(L, l)| = |L(l)| \tag{2}$$

The point of Δ is because we want to be working relatively. Whether one has German on Tuesday or Thursday is a useless fact unqualified: what is more interesting is if, for example, one hasn't had German since the Wednesday before. To put the (poorly notated) definition in English, the sequence $\Delta(L, l)$ is the sequence of differences between the places of two ls in L.

Of course the order of elements in L matters (it is a sequence, after all) but the order of elements in Δ doesn't matter for the purposes of this PDF. It is simply a sequence because Why Not. Hence, whether we choose L_k to be the $(i+1)$th or $(i-1)$th occurrence is irrelevant, to the point where it's changed twice in the writing of this document.

There is one reconciliation that must be made to make the facts given true and useful: although the timetable given in figure 1 is laid out starting in one corner and ending in the opposite, it is a cyclic pattern. If $i = 1$, then the $(i-1)$th corresponds to the final, and it is as if we count backwards off the beginning of the timetable and continue counting as we go back onto the end and to the last occurrence of l. Since this is essentially a division of the size of L by its ls, this gives us

$$\sum_{d \in \Delta(L,l)} d = |L| \tag{3}$$

The next definition we give is a simplification of $\Delta(L, l)$:

$$\Delta(L) = \bigcup_{l \in \mathbb{L}} \Delta(L, l) (:= \Delta) \tag{4}$$

$$\text{thus } |\Delta| = |L| \tag{5}$$

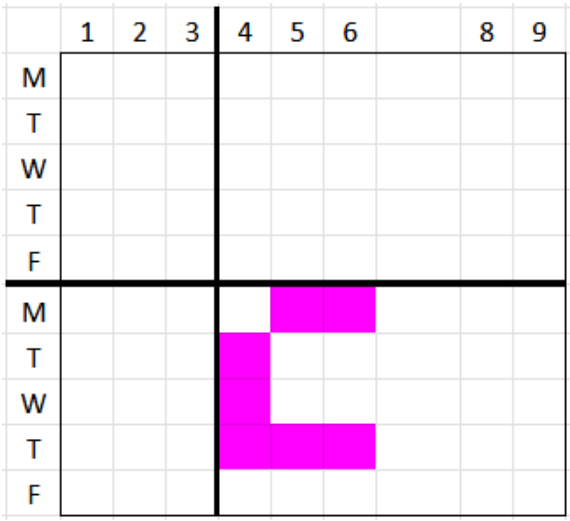

Figure 2: My 2024-25 timetable but with only teacher “F”

This could just as easily be defined from the same angle that was used in (1), but we don’t have the effort to do that presently. The order is irrelevant for the arithmetic in the present work, but for the sake of examples we order them back into the order in which they were in L. The notation of dropping the argument of Δ is simply to make writing slightly quicker and less messy later on.

(5) is also a convenient extension of (2). Is it not lovely when maths is consistent?

Since this has all been rather abstract thus far, let us take a concrete example with a “mini-timetable” made up of a single five-period day that repeats eternally, to aid the reader’s comprehension (see table 1).

L	$\Delta(L,\text{“F”})$	$\Delta(L,\text{“R”})$	$\Delta(L,\text{“C”})$	Δ
F	2			2
R		4		4
R		1		1
F	3			3
C			5	5

Table 1: small-scale examples of Δ

When it comes to applying these methods to a full-length timetable, there are certain decisions that have to be made one way or another. For the sake of usefulness in the present work, we say the following: when one day ends, the next begins immediately (that is to say, the difference between Period 8 and Period 9 on a Tuesday is the same as the difference between Period 9 on a Tuesday and Period 1 the following Wednesday); lunchtimes, break times and weekends are completely ignored (that is to say, the aforementioned difference is also the difference between Period 9 on a Friday and Period 1 on the following Monday, and between Period 6 and Period 8 on any same day). To some extent, we make these points for the sake of simplicity when it comes to interpreting the results later down the line.

3 Working It Out

Now let us think about defining these three terms that I used no more than qualitatively at the beginning of the term, here in reference to my lessons with one of my German teachers, denoted with “F” (see figure 2). Beginning with “Number” because that is trivial to define, we quickly see that some units would help us. The Système international d’unités can’t help us much here: there are a lot of seconds between two lessons. All numbers henceon will instead be given in ”periods” - in reality, forty minutes, but in abstract, any length.

3.1 Number

The "Number" of periods with the teacher in question can be counted by an infant. It is 7. Since our units are well-defined (lessons certainly feel like they're moving slowly enough not to worry about relativistic effects...), we count a "double-" or "triple-period" as 2 or 3 separate periods respectively, which is an attitude that will help us later. By (2),

$$N(\Delta(L,l)) = |\Delta(L,l)| = |L(l)| \tag{6}$$

3.2 Spread

It is no doubt that the lessons of the teacher highlighted in figure 2 are "poorly spread". What do we mean when we say this phrase though? In this phrase, we're referring to the large timeframes that have absolutely no lessons with this teacher: a frame of 55 periods, no less, which is over two-thirds of the entire timetable. A "well-spread" distribution of lessons with this teacher would have these gaps minimised.

Thus, we want to perform some sort of iterative calculation on our sequence of gaps $\Delta(L,\text{"F"})$ that puts weight on the longer gaps. We could attempt a sum of Δ since the longer gaps would thusly account for a larger proportion of the sum. However, the reader who reads can read that the sum of Δ, as per (3) is just the total number of periods, $|L|$.

Instead, we want to put an extra weight on our longer gaps. Luckily, larger values have an in-built tool to weight them favourably: themselves. We can weigh the sum by the elements themselves, which is known to the common mathematician as "squaring" it. Thus, we're dealing with some variation of

$$\sum_{d\in\Delta(L,l)} d^2 \tag{7}$$

However, the same readers who happen to be reading closely will realise that this is almost what we wanted: a poorly-spread distribution will have larger gaps, and thus this sum will be disproportionately *larger*. Thus, the more useful definition to take is a reciprocal of (7).

There is one final amendment to this definition: currently, a sequence of twenty lessons distributed evenly throughout the week gets an unfavourable score in comparison to a sequence of two lessons distributed similarly evenly, by simple virtue of there being more lessons. This is an explicable advantage (it's unfair to expect the latter sequence to populate the timetable as evenly as the former), and so a combatable one: we divide through by the number of lessons (which, by (2) is also the number of gaps). This finally gives way to our definition of "spread":

$$S(\Delta(L,l)) = \frac{1}{|\Delta|}\left(\sum_{d\in\Delta(L,l)} d^2\right)^{-1} \tag{8}$$

3.3 Clump

Let us now whisk through the same process for the word "clump". This is in fact a poor choice of word as one's immediate thought may be that "clump" is the antithesis to "spread", as you can see that the lessons highlighted in figure 2 are "not very spread" and "very clumped". However, let us treat it as a separate phenomenon and instead we will find that a distribution of lessons can in fact be both "well-spread" and "clumped" or "poorly-spread" and "not clumped".

When we say that the lessons in figure 2 are "clumped", we are this time saying that the majority are in close proximity to one another. Specifically, we refer to the fact that most distances in Δ are very very small. In the case of the "triple-period" on Thursday, there are two elements of $\Delta(L,\text{"F"})$ that are just 1! (Interpret that as either surprise or factorial, because it is, by chance, unambiguous.) In a similar fashion to how we defined "spread", we use the reciprocal squares to place weight on these shorter values. This also means that we no longer have to deal with larger values mapping to "less clumped" distributions, as reciprocating our smaller distances gives us a larger "clump". This gives us the following definition of "clump":

$$C(\Delta(L,l)) = \frac{1}{|\Delta|}\sum_{d\in\Delta(L,l)} d^{-2} \tag{9}$$

L	$\Delta(L,$"F"$)$	$\Delta(L,$"R"$)$	$\Delta(L,$"C"$)$	Δ
F	2			2
R		4		4
R		1		1
F	3			3
C			5	5

Table 2: table 1 repeated for your convenience

$l \in \mathbb{L}$	$N(\Delta(L,l))$	$\sum_{d\in\Delta(L,l)} d^2$	$S(\Delta(L,l))$	$\sum_{d\in\Delta(L,l)} d^{-2}$	$C(\Delta(L,l))$
F	2	13	$\frac{1}{26}$	$\frac{13}{36}$	$\frac{13}{72}$
R	2	17	$\frac{1}{34}$	$\frac{17}{16}$	$\frac{17}{32}$
C	1	25	$\frac{1}{25}$	$\frac{1}{25}$	$\frac{1}{25}$

Table 3: small-scale examples of N, S and C

4 Interpreting the Values

4.1 Relatively

Let us use the same "mini-timetable" as in table 1 to give some sample values for C and S, as it's easy to see where values come from on such a small scale: see table 3.

Perusing table 3 and comparing to what we would qualitatively say about the "mini-timetable" (reproduced in table 2), we can see that it's pretty spot-on: teacher "C" has the most spread-out lessons, but teacher "F" is very close behind, for example, while "R"'s "clumpedness" is definitely coming to light. There is a small failure of using a small sample size, in that the scaling suffers a little, and small discrepancies arisen by the number of lessons don't hide themselves. Ignoring that, we can see how this metric has potential.

4.2 (Somewhat) Absolutely

However, this is only good for comparing different lessons within one timetable. Since we have a Δ-sequence for the entire timetable, performing "NCS-analysis" on that is, in its current state, not very useful as we'd have nothing to compare it against, or we may find ourselves comparing apples to pears. We presently briefly restate our N, S and C functions slightly more neatly and using the Δ-sequence of the whole of L. It is now that we remind the reader that "Δ" unqualified represents this complete sequence, $\Delta(L)$.

$$C(\Delta) = \frac{1}{N(\Delta)} \sum_{d\in\Delta} d^{-2} \tag{10}$$

$$N(\Delta) = \sum_{d\in\Delta} d^{0} \tag{11}$$

$$S(\Delta) = \frac{1}{N(\Delta)} \Big(\sum_{d\in\Delta} d^{2} \Big)^{-1} \tag{12}$$

Unfortunately for the reader, we're unable to align the equations exactly as we'd like while still giving them different reference numbers. The reader will have to cope. However, this formatting does demonstrate the relationship between the three functions: if one ignores the reciprocating adjustment we to S, they're in a neat series. The reader with a particularly tenacious memory will remember that (3) fits into this progression, but this is just at a superficial level as (3) references $\Delta(L,l)$ rather than Δ. However, we will return to this insight later.

(11) can be rewritten as the following, and thus by (5) as

$$N(\Delta) = |\Delta| = |L|$$

Using the same order of thinking as with section 3, we begin with regularising values of S. S is formed from the sum of squares: if we take a step back, we have the average of squares,

$$\frac{1}{|\Delta|} \sum_{d \in \Delta} d^2 \tag{13}$$

The average of squares, which we henceon notate as $\hat{\mu}_S^2$ to highlight that it's weird, can be square-rooted to give a plain "average": we're simply undoing what we did to construct $\hat{\mu}_S^2$ to get an ordinary average of gaps but "skewed" in favour of longer gaps, as per the processes in section 3. This gives

$$\hat{\mu}_S = \sqrt{\frac{\sum_{d \in \Delta} d^2}{|\Delta|}} \tag{14}$$

And similarly,

$$\hat{\mu}_C = \sqrt{\frac{|\Delta|}{\sum_{d \in \Delta} d^{-2}}} \tag{15}$$

Simple dimensional analysis reveals that these retain the same units as the elements of Δ to begin with: thus, we can interpret the values of $\hat{\mu}_C$ and $\hat{\mu}_S$ to be the average number of periods between a lesson and the next time that lesson occurs, weighted in favour of clump and spread respectively.

For example, the "mini-timetable" in table 2 yields the following values:

$$\hat{\mu}_C = 1.848\ldots$$
$$\hat{\mu}_S = 3.317\ldots$$

It was roughly at this point in the present article's draft that a friend pointed out the similarity between the formula for "spread-weighted mean" $\hat{\mu}_S$ in (14) and the formula for "standard deviation" σ of a dataset. As far as we are concerned in the present work, this is purely coincidental. However, any further insight from the reader is welcomed.

4.3 (Somewhat) Better

The reader is forgiven if the end of the last subsection felt anticlimactic. The numbers and methods outlined therein were developed after the methods in the proceeding subsection herein, but were presented first due to slightly more elegance and slightly less complexity. We do not see much use in $\hat{\mu}_{S|C}$, hence why they were notated with a hat (such that we can justify ignoring them because they are ugly).

Instead, we take a step back to $\Delta(L, l)$ and attempt to use the same means to evaluate an equivalent of $\hat{\mu}_{S|C}$ for each l. Most of our thinking work has, however, been done in the preceding subsection, with just one major addition.

The perceptive reader with an English Literature qualification will recognise that the sum from (3) remains unused. Now that we are dealing with $\Delta(L, l)$ we can complete the pattern if not only because it is pleasing to typeset:

$$\sum_{d \in \Delta(L,l)} d^{-2} \qquad \sum_{d \in \Delta(L,l)} d^{-1} \qquad \sum_{d \in \Delta(L,l)} d^{0} \qquad \sum_{d \in \Delta(L,l)} d^{1} \qquad \sum_{d \in \Delta(L,l)} d^{2} \tag{16}$$

(16.-2) and (16.2) are used in the definitions for C and S respectively, while (16.0) is trivially $N(\Delta(L, l)) = |\Delta(L, l)|$. If anyone would like to come forward with a use-case or theoretical explanation of (16.-1), they would be more than welcome.

We use (16.1) to define the conventional "mean" of $\Delta(L, l)$. This is hopefully not a new idea to the reader, although we do denote it with a hat to continue the idea that it is the "imperfect mean" or the "mean" of $\Delta(L, l)$ for some l:

$$\hat{\mu} = \frac{\sum_{d \in \Delta(L,l)} d}{|\Delta(L, l)|} \tag{17}$$

This gives us something to compare $\hat{\mu}_{C|S}$ against. Rather than succumb ourselves to the "mini-timetable" again, this property is demonstrated with the timetable represented in figure 1: see figure 4.

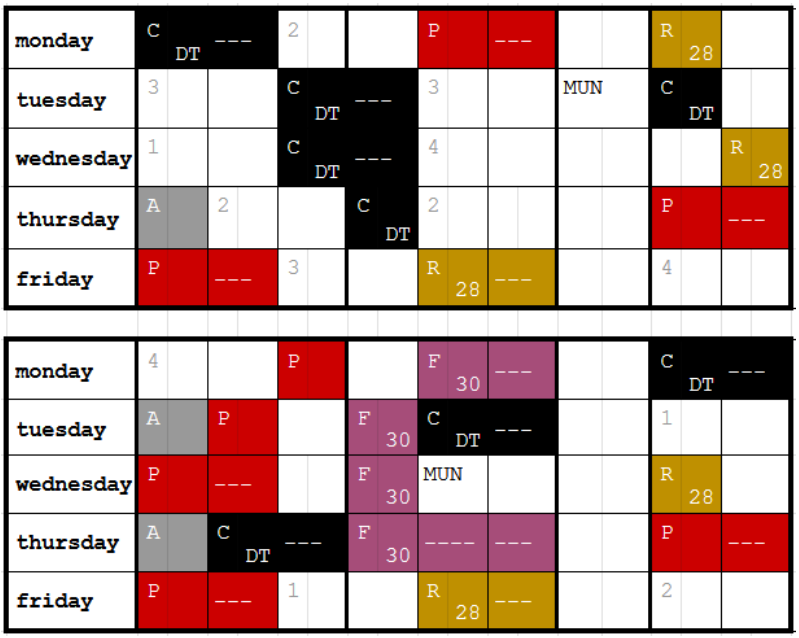

Figure 3: My 2024-25 timetable, repeated from figure 1 for your convenience

	A	B	C	D	E	F	G	H	I	J	K	L
1		-2	-1	0	1	2	C	N	S	μ_c	μ	μ_s
2	c	6.25	7.22	14	80	902	0.45	14	0.000079	1.496	5.714	8.027
3	p	8.06	8.60	14	80	1102	0.58	14	0.000065	1.317	5.714	8.872
4	r	2.02	2.35	7	80	1362	0.29	7	0.000105	1.858	11.429	13.949
5	f	3.05	3.43	7	80	3192	0.44	7	0.000045	1.513	11.429	21.354
6	b	1.57	2.62	9	80	1252	0.18	9	0.000089	2.39	8.889	11.795
7	1	3.00	3.12	6	80	2174	0.5	6	0.000077	1.413	13.333	19.035
8	2	4.36	4.89	8	80	2902	0.55	8	0.000043	1.354	10	19.046
9	3	3.11	3.39	6	80	3262	0.52	6	0.000051	1.388	13.333	23.317
10	4	4.00	4.07	6	80	3774	0.67	6	0.000044	1.224	13.333	25.08

Figure 4: numerical examples of variables given thus far performed on the timetable in figure 1

A brief explanation of the layout of figure 4 may be in order. The rows are headed by the name of each block ("b" meaning "blank" or "unlabelled", and "a" going unrepresented due to there only being three "a" periods in the fortnight), corresponding (almost) exhaustively to the values of $l \in \mathbb{L}$. Columns B through F correspond to the five expressions in (16) respectively, while columns G, H and I correspond to $C(\Delta(L, l))$, $N(\Delta(L, l))$ and $S(\Delta(L, l))$ respectively. Columns J, K and L are, as one would expect, $\hat{\mu}_C$, $\hat{\mu}$ and $\hat{\mu}_S$, despite the lack of hat in the column heading (which proved confusing later down the line). Columns G, I, J, K and L have each been rounded to a number of decimal places for brevity and legibility.

Some numbers suddenly make sense when looking at a table. For example,

$$\forall \mathfrak{L} \in \mathbb{L}, \sum_{d \in \Delta(L,\mathfrak{L})} d^1 = |\Delta| = \sum_{l \in \mathbb{L}} N(\Delta(L,l)) \tag{18}$$

We can also begin making observations similar to those we made looking at the sample tables 2 and 3. For example, all blocks appear mostly in "double-periods", with the exception of "b", which appears (by its nature) mostly in "singles": this is reflected in its $\hat{\mu}_C$ being closer to its $\hat{\mu}$ than any other l. The infamously poorly-spaced-out "f" lessons had the highest $\hat{\mu}_S/\hat{\mu}$ ratio and the lowest $\hat{\mu}_C/\hat{\mu}$ ratio of any lesson blocks, even though the latter was by a smaller margin. This last fact is down to the fact that although the lessons are highly clumped, it's relatively not to a surprising extent. At my year group, the overwhelming majority of lessons are "doubles" (for better or for worse) (which is also reflected in the artificial segmentation of free periods), and "F" has just a good a good proportion of "singles" and "doubles" as any other block. Perhaps the choice to treat these "doubles" as two "singles" is limiting the scope of this specific implementation.

Now to evaluate the table as a whole: instead of combining the blocks when they were in sequence form (as we did with Δ) we combine their respective $\hat{\mu}_{C|S}$. A weighted average has exactly the properties desirable. Although it produces a clunky formula for our "clump" and "spread", it cancels out for the "mean" due to the relation in (18).

$$\mu = |\mathbb{L}| \tag{19}$$

(We briefly interject to alternatively offer the following ugly definition of μ, which doesn't require the proceedings to be calculated for *all* $l \in \mathbb{L}$, as was done in the example where "A" was omitted in the calculations. The discrepancies are small either way, but the following expression gives a worse definition of "mean" and a better comparison point for $\mu_{C|S}$.)

$$\mathbb{M} \subset \mathbb{L} : \mu = \frac{|L| \sum_{l \in \mathbb{M}} 1}{\sum_{l \in \mathbb{M}} N(\Delta(L,l))}$$

Thus, in the last few proceeding definitions, any $\mathbb{L}$ can be swapped for some $\mathbb{M}$ (and thus $|L|$ for $|L(\mathbb{L})|$) with limited loss of accuracy as long as $\mathbb{M} \approx \mathbb{L}$ and/or $L(\mathbb{M}) \approx L$.

$$\mu_C = \frac{\sum_{l \in \mathbb{L}} \hat{\mu}_{C(\Delta(L,l))} N(\Delta(L,l))}{|L|} \tag{20}$$

$$\mu_S = \frac{\sum_{l \in \mathbb{L}} \hat{\mu}_{S(\Delta(L,l))} N(\Delta(L,l))}{|L|} \tag{21}$$

If the reader feels the compulsion to expand (20) and (21) with the square roots given in (15) and (14) respectively, they must first remember to change the sequence in question from Δ to $\Delta(L,l)$ and then to regret their life choices.

What finally remains to round off is to give the values that we have been working towards for the duration of this subsection that has frankly gone on too long. With reference to the L reproduced in figure 3 and using $\mathbb{M} = \mathbb{L} \setminus \{\text{"A"}\}$,

$$\mu = 9.35\ldots \qquad \mu_C = 1.55\ldots \qquad \mu_S = 14.89\ldots$$

5 Sample Implementations

We presently present some exemplars on how this "NCS Analysis" could be used to make observations on various topics.

The majority of these examples take data from the author's personal collection: such personalised data are not to be taken seriously. In fact, this article is written as part of a series of articles to celebrate the end of the fourth year of "a fact a day", as the annual tradition of doing primary research for once. This year's New Year's Eve "fact of the day" is a collection of statistical analyses of various facts about the author themself, of which this section is forming a component.

5.1 Timetable *Verschlimmbesserung*

Verschlimmbesserung (literally something along the lines of "worse-bettering") is a German word that can only be translated roughly to "disimprovement" or "enshittification", but instead it refers specifically to when attempts at improvement just make it worse. This is a phenomenon observed particularly among the bureaucratisation and digitalisation of society, which is exactly what's been happening to the timetabling process at the author's school. Timetabling over a thousand timetables in over a hundred rooms for 90 periods a fortnight is understandably a formidable task, but recently the way it's been done has changed and with it, many have reported a worsening of timetables. This is to what I attributed my surprising distribution of German lessons, as described at the beginning of section 1. However, we presently outline that this is not the case:

Timetables from three years were serialised and "NCS Analysed" with exactly the same methods as demonstrated on my timetable of the academic year 2024-2025 in section 4.3. The data from 2024-2025 is the data as given previously, and the data from 2023-2024 and 2022-2023 are both timetables also from the same year group as the first given one. A small amount of conforming was done to make the timetables as comparable as possible, as different people have different numbers of teachers and similar discrepancies can exist. (Further Mathematics, for example, is taught by four teachers simultaneously, whereas Physics only by one.) The results are surprising (all values rounded to three significant figures) and reproduced in table 4.

Timetable of	μ	μ_C	μ_C/μ	μ_S	μ_S/μ
'22-23	11.9	1.50	0.127	21.0	1.77
'23-24	10.7	1.45	0.135	18.8	1.75
'24-25	9.35	1.55	0.166	14.9	1.59

Table 4: "NCS Analysis" on timetabling development over three years

These results provide counterexample to the anecdote given in the introduction: that the timetabling has not been getting markedly worse in its distribution of lessons, perhaps even better. This is by no means conclusive, as analysis was performed on an extremely small sample size over an extremely short developmental timeframe. Moreover, this is just one descriptive measurement: one could argue that the "triple period" the German class this year received is enough to describe the German class's timetable this year as *verschlimmbessert*. Similarly, the number of clashes that have been timetabled in this year has gone up from more-or-less zero to a non-negligible number. However, by the measures of "clump" and "spread", there are certainly differences between these almost-randomly chosen timetables from three different points in time.

5.2 Focussing

I keep a log of how "focussed" I am for every activity that I do. The data are integers between 1 and 5 and corresponds purely to the ability to focus, dubbed "focability" (so not necessarily on one thing, as this is logged by a separate variable, "monobility"). Each datum, as we use it in this subsection, comprises a starting timestamp, an ending timestamp, an activity label, and a set of values in [5] (for example "focability" and "monobility"). For this analysis, we simply took the "focability" data as a one-dimensional array and ignored the timestamps. This does mean that the results are limited in their scope because there may be unaccounted-for biases when we ignore all durations. (Some "focability" data arguably should've held more weight, as they could last 12 minutes or 12 hours.)

"Focability" data were taken from the Master Log (called "O'Brien", for those that know it well) from a 130-day period between early August and early December 2024, forming a total of 1308 values. To conform the data to our processes, L is the "focability" sequence and $\mathbb{L}$ is the set of integers [5]. Unlike the implementations in sections 3 and 4, where the timetables "looped round", our sequence is now in a completely linear shape. This means that either the first or last instance of each $l \in \mathbb{L}$ cannot be mapped to an element in $\Delta(L, l)$ and thus $|\Delta(L, l)| = |L(l) - 1|$ and by its definition in (4), (5) changes to $|\Delta| = |L| - |\mathbb{L}|$. These changes are not given their own line because they are entirely ignorable facts, given thought-out notation.

Before performing the analysis, it may prove useful to speculate some hypotheses on what sort of results are informative. The threshold used for inputting a 1 or 5 (the most extreme values) are more stringent than thresholds between other numbers, as it's more remarkable when these inputs

are given (as the inputter must be in a rut of un"focability" or a glut of high "focability"). This is useful only to a certain extent: too few 1s or 5s gives little sample size on other information regarding gluts and ruts. We can check the quality of these personal and entirely subjective thresholds by checking whether $N(\Delta(L, l))$ is roughly normally distributed over $l \in \mathbb{L}$.

There is another fact of the nature of inputting that could present itself herein. The data are recorded in a spreadsheet by a computer program that takes a numerical input on command from a panel and appends it to the list. However, if the numerical input is empty but the timestamp inputs are not, then the input is recorded as the same l as the previous value. This is for ease of inputting larger quantities of data, but there could also be the fear that values become "sticky" (where the inputter is likelier to leave the "default" value than change the input, especially when the "focability" is low) and thus not truly reflective of the real-world data. This is difficult to distinguish between ordinary time taken for "focability" to change (humans are, after all, slowly fluctuating beings), but a remarkably high "clump" across the board could point towards this "stickiness" being an issue.

Moreover, as just mentioned, the rate of fluctuation and variation and similar attributions will be shown in the "spread" and "clump" of the analysis, helping to strategically quantify the scale of fluctuation of "focability" and possibly aiding with time-planning in the future. (For example, "clump" holds information on the duration of how long one stays in a particular state, so this could help interleave high-focus tasks with low-focus tasks with the right timings.)

"NCS analysis" was performed on Δ first directly, as in section 4.2 with the $\hat{\mu}$ process, and subsequently on $\Delta L, l$, giving the final μ as in section 4.3. The information given by thus process is then compared and contrasted.

With respect to Δ in its entirety,

$$\hat{\mu}_C = 1.78\ldots \qquad \hat{\mu} = 5 \qquad \hat{\mu}_S = 9.09\ldots$$

Table 5 holds the breakdown of values with respect to each $\Delta(L, l)$. Values are rounded to three significant figures, with the exception of values for N.

l	$N(\Delta(L,l))$	$\hat{\mu}_C$	$\hat{\mu}$	$\hat{\mu}_S$	$\hat{\mu}_C/\hat{\mu}$	$\hat{\mu}_S/\hat{\mu}$
1	108	2.09	12.1	20.7	0.173	1.72
2	369	1.64	3.53	4.73	0.463	1.34
3	288	1.72	3.36	4.32	0.512	1.29
4	312	1.82	4.17	6.62	0.437	1.32
5	125	2.34	10.4	17.0	0.225	1.64

Table 5: values for "NCS Analysis" on "focability" data, broken down by $\mathbb{L}$

Completing the process from section 4.3 using the derived data in table 5,

$$\mu_C = 1.81\ldots \qquad \mu = 5 \qquad \mu_S = 7.31\ldots$$

Immediately it is no surprise that $\hat{\mu} = \mu$ and that $\hat{\mu}_C \approx \mu_C$, although we must confess that we did not expect $\hat{\mu}_S$ to differ so greatly from μ_S, for which an intuitive explanation cannot be given. It's an…exercise for the reader.

With regards to the "stickiness" predicted earlier, μ_C is certainly particularly low (corresponding to high "clump"). To put this in words, the average time between one state of "focability" and the next point in time when that state occurred, weighted in favour of "clump", is less than two activities. However, looking at table 5, we speculate that this is not due to "stickiness" of inputting, but rather "latency" or "drag" in the natural fluctuations of the state of the human mind. This is because the greatest geometric deviation of $\hat{\mu}_C$ occurred for $l = 1$ and $l = 5$, the most "drastic" inputs and thus the inputs on which one would expect to find the least "input stickiness". Instead, this points to the "dragging" of "ruts" and "gluts" being the primary factor of an unexpectedly low μ_C and $\hat{\mu}_C$. This is a positive affirmation that the collection of data is of sufficiently high quality to reflect the real world, and that it may be a good capture of an inherently subjective quality.

The discussion in the paragraph above was notionally on the value of μ_C, which was very similar to $\hat{\mu}_C$. The primary difference between these two values was that μ_C took around half an hour of spreadsheeting by an expert spreadsheeter to evaluate, whereas $\hat{\mu}_C$ took around five minutes. This is of course ignoring the processes involved in both paths (e.g. collecting, selecting and conforming the data) but there is undoubtedly greater levels of thinking and computation required to perform the "Better" method. Thus, $\hat{\mu}_{C|S}$ is presented as an approximation of $\mu_{C|S}$.

There is however still merit to the longer process. Having more steps involved makes it more laborious, but equally more useful: the intermediate steps (values of $\hat{\mu}_C$ for each $\Delta(L, l)$ in this case) proved useful for resolving the aforementioned discussion.

5.3 Socks

Another beautiful datalog in my archives is a log of my outfit each day. In this log, each day a "top", "overtop", "pair of bottoms" and "pair of socks" forms a datum. Before we perform an "NCS Analysis", if the reader would indulge us in validating our nerd credentials with a fact: by careful design and planning of day-to-day outfits, no two data since 2023-05-20 are the same. That is to say, no two "outfits" have been the same, to date and for the foreseeable future. This is less impressive when one considers that the majority heavy-lifting is done by the socks, of which there are at least 40 listed pairs and at least 20 pairs in regular use - one could argue that they form an underwhelming component of the "outfit", but this is not the point.

Due to the large number of socks (33 distinct elements $l \in \mathbb{L}$, as of before Christmas 2024 (one can understand that this number increases rapidly every Christmas)), we discuss the limitations of "NCS Analysis" thereon.

One fact that was regrettably not given more emphasis in the previous section is that with a linear log (as opposed to a circular timetable), the first (or last, depending on your choice of k in (1)) occurrence of each l in L can't be mapped meaningfully to a value in Δ. Hence it is omitted, and thus, as discussed previously, by (4), (2) becomes $|\Delta| = |L| - |\mathbb{L}|$. Previously, this posed no problem as $|\mathbb{L}|$ was negligibly small in comparison to $|L|$, but here it pushes the value of N in an unaccountable direction by around one sixteenth, immediately limiting the accuracy of our results thereto. Multiple pairs of socks were only worn once in this one-and-a-half-year period, and thus have no effect on the result - perhaps this is fair or unfair.

Due to the large size of $\mathbb{L}$, the "approximation" method only was used (the method notated with a hat and dealing with Δ in its entirety). To begin,

$$|L| = 588 \qquad |\mathbb{L}| = 33 \qquad |\Delta| = 555 \qquad \hat{\mu} = 33$$

A sample of the data (just 49 entries, thus representing less than a tenth of the whole dataset) is given in table 6 such that the reader can understand at a superficial level what sort of data is being dealt with. For the reader's interest, measures of average and spread for the entirety of Δ are also given, to highlight its peculiar nature. Obviously Δ is strictly positive. The most notable parts are the high standard deviation, and the disparity between non-weighted mean and the value for $\hat{\mu}(= \mu)$ above.

$$\text{mean} = 18.6\ldots \qquad \text{stdev} = 28.3\ldots \qquad \text{mode} = 6 \qquad \text{median} = 12$$

The following results are derived:

$$\hat{\mu}_C = 4.28\ldots \qquad \hat{\mu} = 33 \qquad \hat{\mu}_S = 33.8\ldots$$

$$\frac{\hat{\mu}_C}{\hat{\mu}} = 0.1298\ldots \qquad \frac{\hat{\mu}_S}{\hat{\mu}} = 1.025\ldots$$

The tiny disparity between $\hat{\mu}_S$ and $\hat{\mu}$ comes from an incredible high value for "spread": by and large, most socks were worn remarkably evenly throughout the timeframe. We can vouch anecdotally that this is not true for a non-negligible number of pairs of socks, which were worn for a small sub-period and then not again for a long time (for example if they became hole-ridden and entered the long queue to be darned). This exposes two facts about the processes at play:

Firstly, that removing the circular nature of it means that pairs of socks that were poorly spread to the point that they never appeared again after a certain point do not have this accounted for.

5	9	7	12	23	15	88
9	17	11	106	11	9	6
13	8	5	6	14	8	11
4	10	7	18	6	10	9
2	15	15	1	6	26	16
173	4	4	5	1	1	28
9	1	6	4	40	19	1

Table 6: the last 49 elements of the sequence Δ from the "socks log"

.

That is to say, there are long periods, sometimes spanning the entire timeframe of the data, where there ought to be a "distance" but due to a lack of element in Δ this is not accounted for in "spread" calculations. This may be desirable, but we believe that this could be improved upon. Possibly treating the linear log as a circular log ("patching it up" from the end to the beginning) inserts these periods back in in a neat manner, but may also give rise to less clear problems that cannot easily be seen in the results at 11pm at night. For example (and this is pure speculation), socks that were only acquired into the log at some point during the timeframe of recording might not want to have that initial "lead time" to be counted towards their value for $\hat{\mu}_S$ as this would be "unfair". The optimally-spread scenario in this case would involve it being locally well-spread but not globally as the new sock cannot be spread into the period before which it was adopted, obviously.

Secondly, the socks with a poorer spread, by virtue of them having become holey or unpreferred, also have a lower number and thus count less significantly towards the value for μ_S and by extension its simulated counterpart $\hat{\mu}_S$. This may be a preferred feature of the analysis, helping to give "fair" treatment to socks that become damaged or forcibly removed. It similarly helps hide the weight of socks whose preferability changed, but then it comes down to the purpose for which one wishes to perform the analysis in the first place, if it is not to gauge the magnitude of changes in preferability and fashion choices.

One alteration to the method that may aid to solve some of these uncertainties could be a time-based regressive analysis. For example, a new $\widetilde{\Delta}(t)$ could be formed from a window of L truncated at t and $t+k$, similar to how moving averages are computed. This exacerbates certain aspects of the problems discussed both in this subsection and elsewhere in the document, but helps to analyse others. The ins-and-outs of how to go about this escape the author at this hour, but the reader is welcome to practice their spreadsheet skills hereon.

5.4 Flashcard Reviews: Bins

Anki is a FOSS program to help memorisation, based on a spaced repetition system (SRS) and active recall. I use this program (and associated programs, e.g. AnkiDroid, AnkiWeb) to help learn vocabulary for language-learning, among other things. One of the core tenets of the SRS is that one uses the app on a daily basis to optimise long-term memorisation. The power of the SRS only works when one is consistent. Thus, it may be that an "NCS Analysis" of the review history of an Anki collection could shed some light, as well as revealing some limitations of the analysis process.

There is a component of Anki's culture that involves one's "streak" (how many consecutive days one has reviews flashcards) and one's "heatmap" (the representation of the density of one's reviews). One example of the heatmap is given below (cosmetics may vary) in figure 5. The brighter the cell, the more reviews were performed on that day. It is generally the goal to have one's reviews as evenly distributed throughout the heatmap as possible, as this benefits one's long-term memory and reduces "buildup" (where one becomes indebted to the rising number of "due" flashcards due to the SRS algorithm). High "clump" (low $\hat{\mu}_C$) could arise from "cramming", while poor "spread" (high $\hat{\mu}_S$) could arise from poor time-management.

In figure 5 it can be seen how the reviews were most reliable during the period of exams (due to a regular morning routine incorporating flashcard reviews) and least reliable during the first term of the current academic year (due to the intensity of schoolwork and the exacerbating effects

of the buildup of due cards).

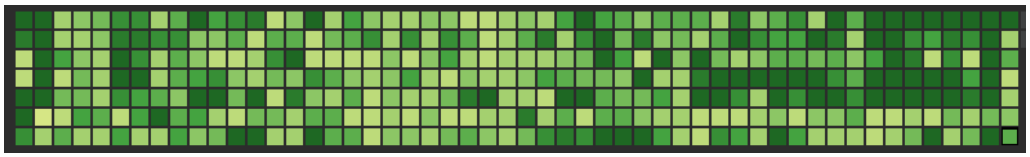

Figure 5: an example of Anki's heatmap, showing data between 2024-1-1 and 2024-12-29

As Anki is completely FOSS, the "revlog" (a database of every review ever) is easily accessible. A list of timestamps were extracted, corresponding to every time a flashcard was reviewed. In this case, as of writing, this is a list of 161,999 points in time, given with millisecond-precision. This was extracted, but needs to be conformed in some manner to L. The process behind the heatmap was simulated: timestamps from this calendar year (84,023 reviews) were collected into 366 bins, corresponding to the days of the year, and the size of each bin formed L.

However, because most values of L is in the order of magnitude of hundreds, the vast majority of elements of L are unique (that is to say, $\mathbb{L}$ is very large) and thus Δ is very small (120). Thus, each day is rounded to the nearest hundred, instead giving $|\mathbb{L}| = 11$. The results of the analysis hence are as follows:

$$\hat{\mu}_C = 1.82\ldots \qquad \hat{\mu} = 11 \qquad \hat{\mu}_S = 18.0\ldots$$

$$\frac{\hat{\mu}_C}{\hat{\mu}} = 0.165\ldots \qquad \frac{\hat{\mu}_S}{\hat{\mu}} = 1.64\ldots$$

Performing the same on the entire history, dating back to November 2022, yields

$$\hat{\mu}_C = 1.71\ldots \qquad \hat{\mu} = 13 \qquad \hat{\mu}_S = 28.7\ldots$$

$$\frac{\hat{\mu}_C}{\hat{\mu}} = 0.132\ldots \qquad \frac{\hat{\mu}_S}{\hat{\mu}} = 2.208\ldots$$

Which can be interpreted to affirm that this year's reviews were a remarkable improvement on the baseline, with significantly better "spread". However, if we take just the last academic term (corresponding roughly to the right-hand third of the heatmap in figure 5, where Mondays-Fridays are notably unpopulated), then we yield

$$\hat{\mu}_C = 1.75\ldots \qquad \hat{\mu} = 8 \qquad \hat{\mu}_S = 9.56\ldots$$

$$\frac{\hat{\mu}_C}{\hat{\mu}} = 0.219\ldots \qquad \frac{\hat{\mu}_S}{\hat{\mu}} = 1.195\ldots$$

This demonstrates an embarrassing value for "clump" but an unexpectedly passable value for "spread". Contrasting with the heatmap, which offers qualitative and hand-wavey results, we can numericise our hunch to a degree, but equally it is evidently limited as we would have expected such a poor pattern to yield worse "spread" than is given.

Thus, the "bins" approach is limited in its usefulness. However, we offer an alternative process in the subsequent subsection:

5.5 Flashcard Reviews: Timestamps

Our data was conformed quite drastically to fit in with the model we developed looking at timetables. However, a completely different model more suited to this type of data could be more prospective.

If we take discard the idea of $\mathbb{L}$, everything suddenly becomes simpler, to the point where we should have begun with this example rather than the timetables. In the case of the timetables, we have points (the beginning of lessons) dotted through a medium (other lessons), and thus the

difference between two lessons is given in terms of other lessons. This was also the case in the two other examples. However, in the case of the Anki review history, we have points (timestamps) dotted through a medium (the passage of time). This gives a much more absolute and sensible derivation for Δ:

$$\Delta_i = L_{i+1} - L_i$$

As previously, there is an offset due to the fact that L_{i+1} runs out and there is one element (the last/first, depending on how one counts it) of L that does not correspond to an element Δ. This is even less of a problem than previously, as this only applies to one L_i, in spite of $|\mathbb{L}| = \infty$ by some twisted interpretation.

As one would expect, this new approach devalidates the definition of "mean" in (19), as this would give $\mu = 1 \forall L$. Instead, we give a more intuitive definition of "mean":

$$\hat{\mu} = \frac{\sum_{d \in \Delta} d}{|\Delta} = \frac{L_{\text{final}} - L_1}{|\Delta|}$$

Performing this updated process on the 2024 dataset of 84,023 datapoints gives

$$\hat{\mu}_C = 3.04\ldots \qquad \hat{\mu} = 371.6\ldots \qquad \hat{\mu}_S = 4348.0\ldots$$

$$\frac{\hat{\mu}_C}{\hat{\mu}} = 0.00819\ldots \qquad \frac{\hat{\mu}_S}{\hat{\mu}} = 11.7\ldots\ldots$$

This at first appears truly shocking, but there is a rational explanation: when one says that one has reviewed n cards in a day, one often can also say that one has reviewed n cards in a small kn time period during that day. For me, that k is usually around 6 seconds. When one revises less in a day, there is still an extraordinarily high clump and poor spread because that revision occurs in a very small period of time in comparison to the day as a whole. This could be seen as fair treatment (perhaps one would learn more if one woke up every ten minutes during the night to review one (1) flashcard) but for this specific case, we perform a "smear":

This is a programmatic method of redistributing reviews throughout a day. For example, if a total of twelve reviews were executed on a date, but within two minutes of one another, the timestamps for these reviews are spread out evenly across the date in question, instead becoming Midnight, 2am, 4am, etc.. This could arguably lose a lot of precision with respect to *when* in the day the the reviews were carried out, but it also more loyally analyses the heatmap reproduced in figure 5. Unfortunately for the author, these subsections exceed the capabilities of spreadsheeting and enter the realm of Shoddy Python Code.

After this "smearing" (which is significantly more computationally intensive, unfortunately) was performed on the 2024 dataset, "NCS Analysis" was performed on the new set of timestamps (all averages rounded to the nearest integer):

$$\hat{\mu}_C = 180 \qquad \hat{\mu} = 374 \qquad \hat{\mu}_S = 1282$$

$$\frac{\hat{\mu}_C}{\hat{\mu}} = 0.6575\ldots \qquad \frac{\hat{\mu}_S}{\hat{\mu}} = 4.674\ldots$$

Since these are all values with units of seconds, they need to be translated to "number of cards per day", which is proportional to the reciprocal of these values. $\hat{h}_{C|S}$ is used to denote these values.

$$\hat{h}_C = 480 \qquad \hat{h} = 231 \qquad \hat{h}_S = 67$$

These are very revealing values and much more useful than those given before. An average of 231 cards per day is good and a stone's throw from my target, and the fact that these stray comfortingly little when weighed by "clump" and "spread" is equally reassuring, as 67 or 480 cards in a day points to no failure, either by neglect or mismanagement.

This year, an active effort was made as a component of my New Year's Resolution to improve the usage of Anki and consistency thereof. It's reassuring that, in spite of the cliff-fall in the last third, good usage was made. To aid these comparisons and evidence the goodness and success of 2024, we give the same analyses for the other two periods also outlined in the previous subsection, with all integers given having been rounded.

For the entire 26-month history:

$$\hat{\mu}_C = 199 \qquad \hat{\mu} = 420 \qquad \hat{\mu}_S = 4043$$

$$\hat{h}_C = 434 \qquad \hat{h} = 206 \qquad \hat{h}_S = 21$$

$$\frac{\hat{\mu}_C}{\hat{\mu}} = 0.474\ldots \qquad \frac{\hat{\mu}_S}{\hat{\mu}} = 9.64\ldots\ldots$$

And for the last third of 2024:

$$\hat{\mu}_C = 217 \qquad \hat{\mu} = 510 \qquad \hat{\mu}_S = 2173$$

$$\hat{h}_C = 398 \qquad \hat{h} = 231 \qquad \hat{h}_S = 40$$

$$\frac{\hat{\mu}_C}{\hat{\mu}} = 0.426\ldots \qquad \frac{\hat{\mu}_S}{\hat{\mu}} = 4.26\ldots\ldots$$

Here's to another year, taking all the successes of 2024 and building upon the failures! May 2025 treat us all better.