Warren Thorngate, Chunyun Ma

Wiggles and Curves : The Analysis of Ordinal Patterns

Problemy Zarządzania 14/2 (2), 160-172

2016

Problemy Zarządzania, vol. 14, nr 2 (60), t. 2: 160–172
ISSN 1644-9584, © Wydział Zarządzania UW
DOI 10.7172/1644-9584.60.10

Wiggles and Curves: The Analysis of Ordinal Patterns

Submitted: 02.07.16 | Accepted: 19.08.16

Warren Thorngate[*], **Chunyun Ma**[**]

Almost all social science data are analysed with variants of the General Linear Model (GLM): regression analyses, analyses of variance, factor analyses, path analyses and the like. However, many interesting and important social phenomena cannot be addressed with the GLM. Ordinal Pattern Analysis (OPA) was developed to examine such excluded phenomena. OPA is a goodness-of-fit procedure for calculating indices of how well a researcher's ordinal predictions match the ordinal properties of data at hand. While the GLM requires raw data to be aggregated across individuals or groups first before being analysed, OPA permits the reverse: Raw data from each individual or group can first be analysed, then aggregated. The reversal reveals what occurs "in general" rather than "on average" – two revelations that often diverge. We illustrate some uses of OPA with simple examples, and provide a computer programme for expediting OPA calculations.

Keywords: statistics, inference, goodness-of-fit, ordinal, pattern.

Analiza Struktur Porządkowych

Nadesłany: 02.07.16 | Zaakceptowany do druku: 19.18.16

Prawie wszystkie dane w naukach społecznych są analizowane przy użyciu wariantów ogólnego modelu liniowego (GLM): analizy regresji, analizy wariancji, analizy czynnikowej, analizy ścieżek i tym podobnych. Jednak wiele ciekawych i ważnych zjawisk społecznych nie daje się przeanalizować przy użyciu GLM. Analiza Struktur Porządkowych (OPA) została opracowana w celu zbadania tych wykluczonych zjawisk. OPA, oparta na kryterium dobroci dopasowania, służy do obliczania wskaźników dopasowania przewidywań badacza do rzeczywistych danych mierzonych na skali porządkowej. Podczas gdy GLM wymaga agregowania surowych danych po obserwacjach lub grupach przed przystąpieniem do analiz, OPA pozwala na działanie odwrotne: surowe dane od każdej osoby lub grupy można najpierw analizować, a następnie agregować. Ten mechanizm odwrócenia ujawnia zjawiska, które występują „zazwyczaj", raczej niż „średnio" – dwa wyniki, które często się różnią. Ilustrujemy niektóre zastosowań OPA na prostych przykładach i udostępniamy program komputerowy do obliczeń OPA.

Słowa kluczowe: statystyka, wnioskowanie, dobroć dopasowania, struktury porządkowe.

JEL: C14, C18

[*] **Warren Thorngate** – prof., Carleton University Ottawa, Ontario, Canada K1S 5B6.

[**] **Chunyun Ma** – mgr, Carleton University Ottawa, Ontario, Canada K1S 5B6.

Correspondence address: Psychology Department, Carleton University, 1125 Colonel By Drive Ottawa, Ontario K1S 5B6, Canada; e-mail: warren.thorngate@carleton.ca.

Ministry of Science and Higher Education
Republic of Poland

The creation of the English-language version of these publications is financed in the framework of contract No. 768/P-DUN/2016 by the Ministry of Science and Higher Education committed to activities aimed at the promotion of education.

Beginning in the 1950s, people who analysed social science data diverged into two camps: Goodness-of-Fitters (Fitters), and Variance-Splitters (Splitters). Fitters wanted to assess how close the predictions of their disciplinary theories or models mimicked the data they collected – how close was their theory-data fit. Most of their models were mathematical, and most of their observations had the useful properties of ratio measurement. Prototypical were mathematical models of perception and learning that generated beautiful predictions of regular wiggles and smooth curves.

Many Fitters tested their predictions with data gathered from laboratory experiments. They plotted the results on the same graph as the predictions, then assessed how closely the results fit the predictions. Some Fitters adapted least-square indicators similar to those employed by Splitters. Rather than praying for statistically significant differences, however, the Fitters prayed for insignificant ones, hoping to conclude that their predictions and observations were insignificantly different, if not quite the same. Other Fitters relied on a more venerable tradition from the physical sciences, ignoring statistical inference tests for what became known as the "eyeball technique" or "inter-ocular-trauma test": assessing fit by whether or not plots of predictions and observations were close enough to hit them between the eyes.

In contrast to Fitters, Splitters spent little time inventing or refining theories and models using concepts from their own disciplines. Instead, they chose to embrace a statistical model and theory of inference that promised to reveal patterns in data, even patterns smothered in sampling and measurement error, without the need of disciplinary concepts. Their choice? The General Linear Model (GLM) grafted onto traditional, Neyman-Pearson (NP) tests of a misleading euphemism: *statistical significance*.

During the past half-dozen decades, the population of Splitters in social science has grown exponentially, while the population of Fitters has sputtered. The reasons have nothing to do with the relative validity of each approach. Indeed, as the literature exposing the numerous limitations of the GLM + traditional inference has expanded, goodness-of-fit appears increasingly as a sensible alternative. But, much like the silly norms of English spelling, the popularity of the GLM+NP style of data analysis has been much influenced by psychological, social and historical factors. Included in are aversions to mathematics and critiques of statistical induction, the misperceived sophistication of omni-variate GLM derivations, norms of publication, generational lags in the teaching of statistics, and the seduction of menu items in commercial statistical software.

Yet, despite its continuing popularity, the limitations of variance-splitting remain. They begin with its fundamental assumption: It is scientifically meaningful to decompose or partition variations in one set of measurements (independent or predictor variables) into those related to variations in other sets of measurements and those related to "error." The assumption is eerily similar to that early astronomers who accepted Ptolemy's assertion that the earth was the centre of the universe, and that heavenly bodies (planets and stars) revolved around the earth in perfect circles or cycles. When observations of these bodies did not match their predicted locations, astronomers invented the concept of an epicycle, a smaller revolution coiling around the perfect circle, then an even smaller second revolution coiling around the coil around the perfect circle...

Thanks to Copernicus, Kepler and others, Ptolemy's assertion eventually sank into the pit of bad ideas, only to be revised by the GLM and its partitioning of variance into the equivalent of cycles and epicycles. Think of the grand mean of a data set as its earth. Main effects (cycles) revolve around the grand mean. First-order interactions (epicycles) revolve around the main effects. Higher-order interactions revolved around the next-lowest ones. And whatever is left is assumed to be random sampling or measurement error. These assumptions make the GLM mathematically beautiful and relatively easy to extend (as some used to do in stats courses before the invention of statistical software). But its beauty seems unmatched by its realism; there is scant evidence beyond astrology to believe people behave according to epicyclic laws.

Other limitations of the GLM have less to do with its assumptions or derivations than with how it is employed. Consider, for example, the tendency among social scientists to believe that whatever they are studying will be revealed as differences in averages or means. One consequence is that differences in the variability or in the shape of sample distributions are either overlooked or treated with contempt. Many social scientists are taught to conduct preliminary tests of variance differences, such as Levene's test or the F-Max test, praying that the tests will not be significant. If they are significant, cursing is permitted, as are transformations to reduce the variance differences and culls of offensive outliers. Could such variance differences be associated with something important, replicable and real? Perhaps. But the answer lies outside the tradition of the GLM usage.

Another limitation comes from social scientists believing that the mean is, well, meaningful or conceptually proper. The statistic called the mean has some admirable mathematical properties when data are normally distributed. Otherwise, the scientific usefulness of the mean rapidly declines. Consider, for example, an experiment in which 1,000 research participants are asked to rate how much they like or dislike Marmite (a brewing yeast extract) before and after tasting it for the first time. Before tasting all 1,000 participants rate Marmite as a 0 on a scale ranging from –3 (hate it) to +3 (love it). After tasting the goop, 500 rate it as +3 and 500 rate it as –3.

Did the tasting make a difference in participants' liking of Marmite? On average, no; the average rating of Marmite after tasting was 0, exactly the same as before tasting it. But in general, yes; 100% of participants

DOI 10.7172/1644-9584.60.10

changed their rating. Which answer is correct? Both. Which answer gives us more insight about people's reaction to Marmite? We vote for what people did in general. What people do in general is frequently different and more meaningful than what they do on average. Most decision researchers learn this after they naively average across the choices of people who use different decision processes, thinking that an "average decision process" will be revealed. It won't, because an average decision process is about as meaningful as an average dinner recipe or computer algorithm.

One more limitation is worth mentioning here. Despite repeated warnings and calls to cease, most social scientists still use significance testing as a filter, a means of deciding which results they will attend to and which they will ignore. The significant ones show up in graphs and discussion sections; researchers pay attention to them as indicants of the success of their research endeavor. Insignificant differences are given short shrift, a dismissive sentence or two at best. Yet significance testing remains subject to the Law of Large Numbers: The bigger the sample is, the more likely a significant difference will be found. It is the paradox of the concept of statistical power. We want the power to detect some significant differences, but we don't want so much power that all differences are statistically significant.

How do we increase statistical power? Either we buy power with assumptions (Coombs, 1968) or we buy it with large samples. Assumptive power is illustrated when we assume normal distributions and independent observations in order to apply parametric tests. Sample-size power is illustrated in the frustrations WSAD Summer School students had interpreting their printouts when analysing sample sizes of 10,000 or more survey respondents. Everything was statistically significant: Every main effect, every interaction, every correlation, and every test for anything else. Leaving many students baffled by what to do next.

What to do? The traditional notion of statistical significance addresses the relationship between samples and their populations. In particular, it gives us indicants of the risks – the alpha errors and beta errors – we face in generalizing from samples to populations. This is often important when the goal of research is to estimate, say, the outcome of an election based on small samples of voters or the chances of a contracting a disease based on small samples of those exposed to it.

But this is only one of two definitions of generalization. The other one, we submit, is far more suited to the kind of social science research most social scientists do – the kind that looks for reasons why some patterns of results occur and others don't. And it brings us back to overlooked statistical practices of Fitters.

Evidential statistics

Most people taking their first statistics course are required to believe two false assertions. The first is that some statistics are better than others; many people believe that means, for example, are better than medians or modes, and that parametric tests are better than their poor, nonparametric cousins. This might be true if "better" refers to mathematical elegance, but it is absurdly false if better refers to what is useful for examining the data at hand (see, for example, Bradley, 1968).

The second false assertion is that statistics come in only two flavours: descriptive and inferential. Statistics come in at least three flavours, each serving an important purpose. *Descriptive statistics* summarize properties of samples. *Inferential statistics* indicate how well our samples can generalize to populations. *Evidential statistics* indicate how well our predictions can generalize to our samples.

We believe that social sciences would be well-served by making more use of evidential statistics, assessing how well the patterns of their predictions match the patterns of data in their samples of observations. How can this happen? Most Fitters live in a world of mathematical models and precise, laboratory measurements that generate ratio-scale data to test numerical predictions. The rest of us don't. In our world, theories and models are verbal. We make greater-than, less-than, rather than numerical, predictions. And most of the data we collect, such as rating-scale data, rarely have more than ordinal properties.

Are there goodness-of-fit tests adapted to these verbal and ordinal realities? In particular, are there statistical procedures for generating indices of how well the ordinal properties of our data fit our ordinal predictions? And can these indices be used to assess the fit of the raw data produced by individuals as well as aggregated data summarized by descriptive statistics?

In a word, yes.

Inspired by Denys Parsons' (1975) ingenious method of encoding musical tunes by the ordinal properties of their adjacent notes (higher, lower, repeat), the first author has for about 40 years indulged in the nerdish hobby of creating ordinal, goodness-of-fit tests. The result: Ordinal Pattern Analysis (OPA), a collection of evidential statistical procedures for quantifying the fit between ordinal predictions and the ordinal properties of observations. OPA generates what we believe are meaningful and useful quantitative indices of these prediction-observation fits. Its procedures can be easily employed to examine data from single individuals as well as from their aggregates. Several examples of the uses of OPA have been published (for example, Thorngate, 1986a, 1986b, 1992; Thorngate & Carroll, 1986; Thorngate & Edmonds, 2013).

Below we give a brief overview of Ordinal Patten Analysis accompanied by a few examples of its use. We offer three warnings before we begin. First,

DOI 10.7172/1644-9584.60.10

OPA is exceedingly simple to understand and easy to calculate – sufficiently so to be dismissed as a toy in comparison to its uber-omni-multivariate cousins. We, of course, do not share this belief. Second, OPA does not generate p-alpha levels needed for significance testing. It is certainly possible to generate estimates of these levels using resampling/bootstrapping techniques. But OPA is designed to provide a quantitative measure of fit, not to generate a test of statistical significance. In this way, OPA is closer to Tukey's (1977) ideas of exploratory data analysis, and to later notions of data mining, than it is to traditional statistical inference. While OPA can be employed to assist in constructing theories, it was designed to assess the validity of predictions derived from theories. This presupposes that theories exist. If you are hoping that OPA will mechanically generate clear theories from a cloud of observations, disappointment will prevail.

Ordinal Pattern Analysis by example

Let us begin with one of the simplest examples of Ordinal Pattern Analysis we can imagine. Suppose we want to assess a theory of musical performance predicting that musicians will perform better during their recital than during their final rehearsal. We ask 12 cellists playing the same piece to record their final rehearsal and their recital performances, then ask a well-known music critic to rate the 24 recordings given to her in random order. When finished, the critic gives us her 24 assessments: scores on a 10-point rating scale ranging from 0 = terrible to 9 = superb. Table 1 shows her assessments.

musician	A	B	C	D	E	F	G	H	I	J	K	L	average
sex	M	F	F	M	F	F	M	F	M	F	F	M	
rehearsal	2	7	3	NA	4	8	1	6	6	9	4	5	5.0
recital	3	8	5	2	5	9	2	7	7	2	5	5	5.0

Tab. 1. Quality of musical performances in rehearsal and recital

We first note that the average ratings for rehearsal and recital are identical = 5.0. So if we performed a within-subject t-test on the rehearsal-recital rating differences, we would find t = 0.0 then conclude that we found no significant difference in the two means and thus no support for the theory.

But let us look again from the perspective of OPA. We cannot test the recital-is-better prediction on musician D because of missing rehearsal data; he forgot to turn on his recorder during rehearsal. But we have rehearsal-recital rating pairs for the 11 remaining musicians. Of these 11, one received the same rating (= 5) for both the rehearsal and the recital. What should we do with ties? The theory predicts recital rating > rehearsal rating, not recital ≥ rehearsal. Still, ties are often ambiguous; they might, for example reflect a judge's measurement error rather than true performance. OPA's tradition is to ignore ties (for justifications, see Thorngate & Edmonds, 2013).

This leaves 10 musicians whose rehearsal and recital ratings can be compared. What are the chances that, if we selected any of these 10 musicians at random, their rehearsal and recital ratings would show the order our theory predicts? We get the answer by counting the number of these musicians with recital > rehearsal (a *hit*), and the number with recital < rehearsal (a *miss*). Nine of the ten showed an ordinal pattern of their ratings that matched the prediction: nine hits. One of the ten showed an ordinal pattern that mismatched the prediction: one hit. So the probability that the prediction would generalize to our ten sets of observations is

$$\text{p(a hit given the theory)} = \text{pHit} = 9/(9+1) = 0.90.$$

Without the theory, we would expect 50% of our guesses to be correct:

$$\text{p(hit given a random guess)} = 5/10 = 0.50.$$

So using the theory has increased our predictive accuracy of the data set from 50% to 90%, an accuracy improvement of 80%. OPA uses an index of fit similar to Kendall's Tau to express this improvement: The Index of Observed Fit (IOF).

$$\text{IOF} = (\#\text{hits} - \#\text{misses})/(\#\ \text{hits} + \#\text{misses}) = (9-1)/(9+1) = 0.80.$$

The index pHit is vaguely similar to the amount of variance accounted for by employing the prediction. The index IOF is vaguely similar to the correlation between predictions and observations.

Is the increase in performance from rehearsal to recital statistically significant? We can in this case easily calculate, with a simple binomial test, the probability of 9 or 10 musicians doing better in a recital if the chance of a better recital is 1/2 (the classic null hypothesis, H_0). This probability is:

$$\begin{aligned} p &= [10!/(9!+1!) * 1/2^9 * 1/2^1] + [10!/10! * 1/2^{10}] \\ &= 9/1024 + 1/1024 \\ &= 0.01, \end{aligned}$$

a statistically significant difference. This illustrates again that what is true on average (no average difference between rehearsal and recital scores) is not always true in general. In more complex research designs we could also estimate this probability using resampling methods.

But why test for significance at all? Recall that OPA is not designed to make inferences from samples to populations. It is designed to assess how well our predictions match samples of observations. If we wanted to assess the reliability of our match, we could collect data from more musicians to see if they replicated our findings.

To end this example, suppose the sex of each musician was also recorded, and suppose we wished to use this information to test the prediction of a second theory: females will receive higher performance scores than will males. As seen in Table 1, our sample of 12 included seven females and five males. How can we test for sex differences in their performances? Let's begin by asking two related questions: If we selected at random one male and one female from the 12 musicians, (1) what is the probability that the female would have a higher rehearsal score than would the male? (2) what is the probability that the female would have a higher recital score than would the male?

These probabilities are easy to calculate. We begin by generating a set of predicted ordered pairs, a *POP set*, for rehearsal performances. Musicians B, C, E, F, H, J, and K are female; A, D, G, I, and L are male. So an expression such as "{[B, C] > [A, D, G]}" means "We predict that female musicians B and C will have a more highly-rated performance than will male musicians A, D and G." The POP set for rehearsal performances is:

{[B, C, E, F, H, J, K] > [A, D, G, I, L]}.

We will also make the same predictions for the recital scores.

Next, we simply count how many of the observation pairs match (hit) or do not match (miss) the predictions. Examples:

Predict	Observe	Hit or miss?
B > A	B = 7, A = 2	hit
B > D	B = 7, D = NA	NA
B > G	B = 7, G = 1	hit
...	...	
C > A	C = 3, A = 2	hit
C > D	C = 3, D = NA	NA
C > G	C = 3, G = 1	hit
C > I	C = 3, I = 6	miss
...		
K > L	K = 4, L = 5	miss

We then count the hits, misses and ties,

total hits = 21
total misses = 6
total ties = 1

And we calculate pHit and IOF

$$\text{pHit} = 21 / (21 + 6) = 0.78$$
$$\text{IOF} = (21 - 6) / (21 + 6) = +0.56.$$

The indices tell us that employing a theory generating the prediction „females receive higher performance scores in rehearsal than do males" leads to 78% correct predictions, a 56% increase over what we would expect by flipping a coin.

We now shift from rehearsal scores to recital scores, repeating the same procedures as above.

Predict	Observe	Hit or miss?	
	B > A	B = 8, A = 3	hit
	B > D	B = 8, D = 2	hit
	B > G	B = 8, G = 2	hit
	...		
	C > A	C = 5, A = 3	hit
	C > D	C = 5, D = 2	hit
	C > G	C = 5, G = 2	hit
	C > I	C = 5, I = 7	miss
	...		
	K > L	K = 5, L = 5	tie

total hits = 23
total misses = 6
total ties = 6
$\text{pHit} = 23 / (23 + 6) = 0.79$
$\text{IOF} = (23 - 6) / (23 + 6) = +0.59.$

Finally, we calculate the combined hits, misses, and ties for both rehearsal and recital performances.

total hits = 21+23 = 44
total misses = 6+6 = 12
total ties = 1+6 = 7
$\text{pHit} = 44 / (44 + 12) = 0.79$
$\text{IOF} = (44 - 12) / (44 + 12) = +0.57.$

From this we conclude that a theory predicting females perform better than males in both rehearsals and recitals increases the accuracy of our predictions from 50% to 79%, a 57% increase in predictive accuracy expected by predicting randomly.

Should we have averaged the rehearsal and recital scores of each musician before analysing their fit to the sex-difference prediction? We could have done so. There is nothing mathematical in Ordinal Pattern Analysis restricting us to analysing only raw performance ratings, and often OPA analyses of averages result in very similar evaluations. Still, OPA makes it possible to analyse raw data, while traditional inference tests do not. This, in turn, allows us to test predictions both person-by-person and situation-by-situation.

OPA and Big Data

The examples above were kept small and simple to introduce a few basic OPA ideas and procedures. But OPA is not only suitable for the small and simple stuff; it is equally useful in analysing large data sets such as those labeled *Big Data*. Consider, for example, the analyses of national samples of social science data such as those collected in the European Social Survey or in frequent public opinion polls conducted in most countries. Many such surveys have samples exceeding 10,000, which, as noted previously, virtually guarantees that all traditional statistical tests will be statistically significant, leading to hundreds of head-scratching conclusions such as "French females under 30 with a university degree in a social science who live in cities with fewer than 5,000 residents earned significantly less money ($t = 89.5$, $p < 0.0001$) in 2013 than did Swiss males over 45 with a trade school diploma who spoke German in a city of over one million."

So what? Finding such patterns is only half of science. The other half is to explain why the patterns occur, and this requires testable theories. Where do testable theories come from? Not from the theory fairy, nor from a theory boutique. At some point they must be invented, and their invention is more likely when it is preceded by keen observation and thought. OPA is no substitute for thought, but it can be useful for assisting keen observation by directing our attention to regularities or patterns in data we have.

One of the most common uses of traditional statistics is to uncover certain kinds of patterns in the data we have, particularly patterns of statistically significant differences in means. But nature does not reveal all her patterns in these mean differences. OPA can often be useful for detecting some of the other patterns, including patterns in raw data produced by individuals over time, and patterns that change within individuals across situations.

Consider one of several ways we can employ OPA to uncover temporal and situational patterns of individuals hiding in large data sets – patterns not revealed in regression lines or differences in averages. Suppose some (fictitious) Omninational Social Survey asked 1,000 people to provide some demographic or personality information, then to rate their happiness each month for 3 years (M1, M2, ..., M36). They indicate their happiness on a common rating scale: extremely unhappy = 0 ... 100 = extremely happy. At the end of our study each person has given us 36 happiness ratings. A sample of our hypothetical data set is shown in Table 2.

Person	Months									
	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**...**	**M35**	**M36**
Alice	16	20	31	10	19	35	41	...	57	55
Brian	59	46	67	48	NA	49	77	...	75	69
Carla	35	37	24	NA	21	21	28	...	14	11
Dan	90	92	80	20	NA	64	62	...	50	40
Elaine	50	70	33	NA	25	30	10	...	15	20
Frank	20	20	25	15	NA	25	30	...	55	60

Tab. 2. A hypothetical sample of monthly happiness ratings

What patterns should we look for? Some might be found in traditional correlations between background information (age, sex, SES, etc.) and the average happiness ratings. There is nothing wrong with this traditional approach, but it is likely incapable of detecting many other patterns. So let us look among the ordinal properties of our data set for additional patterns – patterns revealed either across people or over time.

We begin by searching for time-related patterns across people. Our goal is to tally for all possible 36x35/2 = 630 pairs of months how many people in each pair show an increase, decrease of no change in happiness ratings. Comparing months M1 and M2 in Table 2, we count four people increasing happiness ratings (Alice, Carla, Dan and Elaine), one decreasing (Brian), and one tied (Frank). Comparing M1 and M3 we find that three people increased their rated happiness (Alice, Brian and Frank) while the three others' ratings declined. We continue tallying in this way, comparing M1 with M4, M5, ... M36. Then we start comparing M2 with M3, with M4, M5, M36. Eventually, we finish all 630 pairs by comparing M35 with M36 (Alice, Brian, Carla and Dan decline while Elaine and Frank increase).

We can then add our 630 tallies, calculate pHit and IOF, and examine the probability of an increase or a decrease in happiness ratings for any two time periods, M1 through M36, we choose. But given the wide individual variation in the wiggles and curves of the happiness ratings, we are likely to find that the probability of happiness going up or going down is about 0.50 (IOF about 0.00), suggesting no strong ordinal pattern overall.

Still, there are more data to be mined, so let's zoom in to search for patterns in more detail. Consider a curious aspect of the ratings in Table 2. During months M4 and M5, five of the participants did not record their ratings; otherwise, they did so religiously. What might have caused the interruption? Notice also that the next available rating after M3 went down for all six of the six people (Alice's declined from 31 to 10; Carla's declined from 24 to 21, etc.). This suggests that something might have occurred to all six people at the end of M3 – a general downer – that interrupted their normal activities and suppressed their ratings. And this, in turn, might lead us on a search for a natural disaster, a local tragedy or news item unique to M3 that caused the happiness dip. If one were found, we might then search all 630 paired comparisons for additional dips affecting most people and a similar search for what economists would call exogenous events. Tukey called this detective work.

There's more. Curiosity about the variability of happiness wiggles and curves such as those illustrated in Table 2 might lead us to examine individual differences. Can the people in the survey be clustered or classified according to their happiness trajectories? One way to answer the question is to devise a measure of similarity or dissimilarity between all possible pairs of survey participants, then employ some form of cluster analysis to group them. There are many different indicants of similarity, and they can generate somewhat different clusters. One of these indicants does for rows in Table 2 what our previous 630 month-by-month comparisons did for columns.

How similar or different are Alice and Brian? Use the order of Alice's 36 happiness ratings to predict Brian's, then use pHit as a measure of their similarity. A pHit of 1.00 indicates that the ranks of Alice's and Brian's ratings fluctuate in unison; a pHit of 0.00 indicates their fluctuations of ranks are mirror images of each other. Do the same to determine how similar or different the wiggles of ratings between Alice and Carla, Dan, Elaine, etc., between Brian and Carla, Dan, etc. are and continue for all pairs of people. After all $(1000 \times 999)/2 = 499{,}500$ distances between participants are calculated, submit them to a cluster analysis programme, and examine the resulting dendrogram for more clues about what distinguishes the groups and subgroups shown.

Conclusions

Thus ends our small exposition of Ordinal Pattern Analysis basics. Limits of time and space prevent us from showing more than a fraction of what OPA can be adapted to do. We refer you to references below in case you wish to learn more.

Unfortunately, OPA will not automate insight, nor will it find meaningful patterns in meaningless data, nor will it uncover causation in correlation, balance your bank account, or cook breakfast. Still, it will provide an alternative means of exploring aspects of data that traditional statistical analyses normally ignore. Its use will reduce the chances of addressing the wrong questions about your data – of committing what Mitroff and Featheringham (1974) call an *error of the third kind*.

We encourage you to try the OPA on some of your own data to get a sense of what it might, or might not, do for you. If you wish to use a small computer programme, written in R, to speed your tallies, please contact the senior author for an electronic copy: warren.thorngate@carleton.ca.

References

Bradley, J. (1968). *Distribution-Free Statistical Tests*. Englewood Cliffs, N.J.: Prenctice Hall, Inc.

Coombs, C. (1964). *A theory of data*. New York: Wiley & Sons.

Mitroff, I. & Featheringham, T. (1974). On systematic problem solving and the error of the third kind. *Systems Research and Behavioral Science*, *19*(6), 383–393.

Parsons, D. (1975). *The Directory of Tunes and Musical Themes*. London: S. Brown.

Thorngate, W. (1986a). The production, detection, and explanation of behavioural patterns. In: J. Valsiner (ed.), *The individual subject and scientific psychology* (pp. 71–93). New York: Plenum.

Thorngate, W. (1986b). Ordinal pattern analysis. In: W. Baker, M. Hyland, H. van Rappard & A. Staats (eds), *Current issues in theoretical psychology* (pp. 345–364). Amsterdam: North Holland.

Thorngate, W. (1992). Evidential statistics and the analysis of developmental patterns. In: J. Asendorpf & J. Valsiner (eds), *Stability and change in development: A study of methodological reasoning* (pp. 63–83). Newbury Park, CA: Sage.

Thorngate, W. & Carroll, B. (1986). Ordinal pattern analysis: A method for testing hypotheses about individuals. In: J. Valsiner (ed.), *The individual subject and scientific psychology* (pp. 201–232). New York: Plenum.

Thorngate, W. & Edmonds, B. (2013). Measuring simulation-observation fit: An introduction to ordinal pattern analysis. *Journal of Artificial Societies and Social Simulation*.

Tukey, J. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*(2), 83–91.

Tukey, J. (1977). *Exploratory data analysis*. New York: Addison-Wesley Publishing Company.

DOI 10.7172/1644-9584.60.10