Urban Maths: Putting the Pieces Together

Rob Ashmore CMath CSci FIMA, Dstl

Inspiration

As I get older, I realise that more and more of my youthful ambitions are likely to remain unfulfilled. I was, for example, barely out of my teenage years before it became apparent that a starring role in the England football team was almost certainly beyond me. It took a little longer, but not too much longer, for me to abandon hope of being a national champion at anything. Or, at least that was the case before I came across the British Jigsaw Championship.

In a way that is, to my mind, quintessentially British, this championship is held annually at St Mary's Church in Newmarket. It is part of the Newmarket Jigsaw Festival, which lasts for six days and includes an ever-changing display of completed puzzles, all of which are available for purchase. Last year's championship had three categories: fun, pairs and elite. Setting my sights on the greatest prize, I looked at the winning time for the elite category in 2018. This was 1 h 55 min, for the 1000-piece puzzle *Ye Old Mill*, which is produced by Gibsons.

Whilst I've done some jigsaws in my time, completing a 1000-piece puzzle in less than two hours is comfortably better than my personal best. To try to improve, I could put in lots of training, but, to be honest, that seems like a lot of hard work: increasing age was not the only barrier to international sporting success! Instead, we are going to do some analysis to try to improve my performance. More specifically, we are going to create a simple model of jigsaw puzzling and use it to determine where I should focus my training efforts.

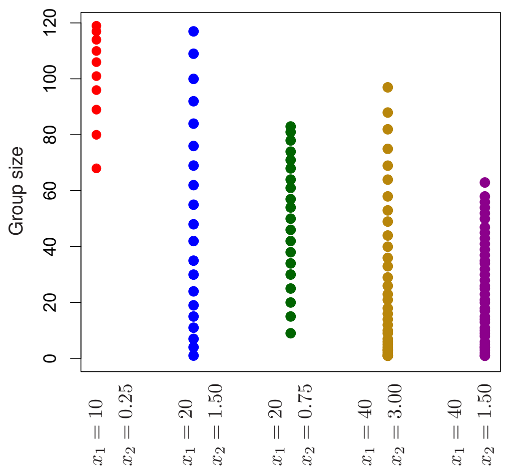

Figure 1: Effect of x_1 and x_2 on group sizes

As best as I can tell, there is no standard analytical model for the completion of a jigsaw puzzle. So, for the purposes of this article we are going to have to make a number of assumptions. Whilst most, if not all, of these assumptions are questionable, the assertion that properly applied mathematics can boost sporting performance is not.

Perspiration

If you'll pardon the pun, the first step in creating a model is to break the problem down into separate pieces. Broadly speaking, there are two aspects to solving a jigsaw puzzle: sorting the pieces into groups and putting pieces from a group into the puzzle.

There is obviously a relationship between the number of groups and the number of pieces in each group. This relationship depends, to a certain extent, on the picture that the complete jigsaw will reveal. To limit the number of parameters involved in our model, we assume that the number of pieces in each of the groups is determined as follows:

1. We create a linearly-spaced vector, spanning from 0 to 1, that contains one more element than the number of groups we will sort into. For example, if there are 10 groups, this vector will be $(0.0, 0.1, 0.2, \ldots, 0.9, 1.0)$. The number of groups is our first parameter.
2. We remove the first element of the vector and raise each remaining item to a power: this power is our second parameter. Continuing our previous example, if the power is 0.25 then the new vector is $(0.562, 0.669, \ldots, 0.974, 1.0)$.
3. We now normalise this vector, and multiply it by the number of pieces in our puzzle (accounting for rounding errors). Elements of the resulting vector are the number of pieces in each of our groups.

Consequently, within our model, we have the following two parameters:

- x_1, which denotes the number of groups;
- x_2, which denotes the power factor.

An indication of how these parameters alter the distribution of group sizes is provided in Figure 1, where the columns represent different cases and, within a column, each point represents a group of a certain size. For example, in the first case (shown in red on the left-hand side of the figure), the number of pieces in a group ranges from 119 to 68, with more groups being towards the upper end of this range. Investigation of Figure 1 shows that our two parameters allow significant variability in how pieces are allocated to groups.

The number of groups also affects both the time taken to place a piece in a group and the probability that a piece is placed in the correct group. With regards to time, we assume this increases as the number of groups increases. Conversely, we assume that the probability a piece is sorted into the correct group decreases as the number of groups increases. Obviously, there are lots of functional forms we could choose to represent these relationships. For our model, we have chosen to use power laws since, firstly, they can be described with only a few parameters and, secondly, they can be used to create curves which initially grow slowly before speeding up, or curves that initially grow quickly before slowing down, or curves that grow at a constant rate.

In our model we have the following parameters, the first two of which relate to time taken, whilst the latter two relate to sorting probability:

- x_3, which is the power describing the increase of sorting time;
- x_4, which is the time taken when we have 40 groups (which is the maximum allowed in our model);
- x_5, which is the power describing the decrease in sorting probability;
- x_6, which is the probability of correctly sorting a piece when we have 40 groups.

An indication of the effect of these parameters is provided in Figure 2, for sorting time (with x_4 being the maximum value on the vertical axis), and Figure 3, for sorting probability (with x_6 being the minimum value on the vertical axis).

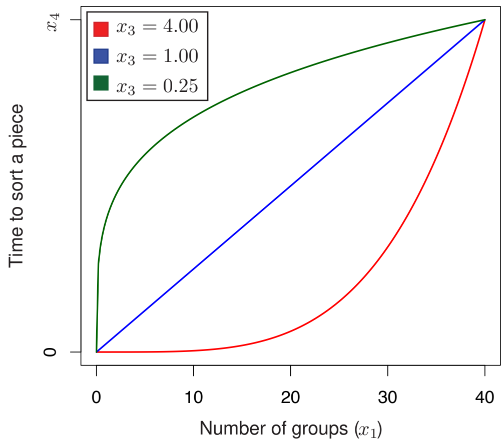

Figure 2: Effect of x_1, x_3 and x_4 on time to sort a piece

The preceding considerations allow us to determine how many pieces are in each group, as well as how many of these have been sorted incorrectly (which, in our model, is just the number of pieces in the group multiplied by the probability a piece has been sorted incorrectly).

We now need to decide how the number of pieces in a group affects the time taken to place those pieces in the puzzle. As with the other parts of our model, there are many functional forms we could choose but, broadly speaking, we expect the time taken to follow an S shape; more specifically, we expect it to be represented by a logistic function. In this case, we assume the minimum value of this function is 1; effectively, all times used in our model are scaled from this value. Consequently, we need three parameters to define our curve:

- x_7, which is the growth rate of the logistic curve;
- x_8, which is the midpoint of the growth phase;
- x_9, which is the time taken to place a piece when there are lots of pieces in the group.

An indication of the effect of these parameters is provided in Figure 4 (with x_9 being the maximum value on the vertical axis).

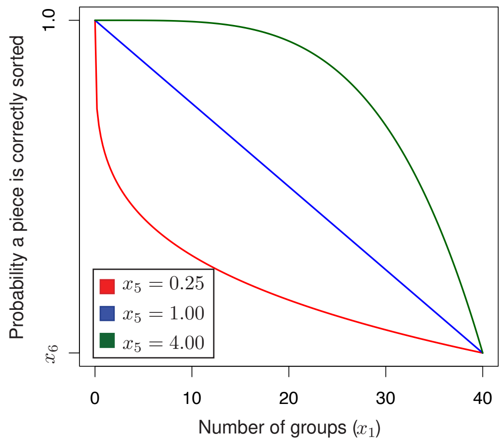

Figure 3: Effect of x_1, x_5 and x_6 on probability of correctly sorting a piece

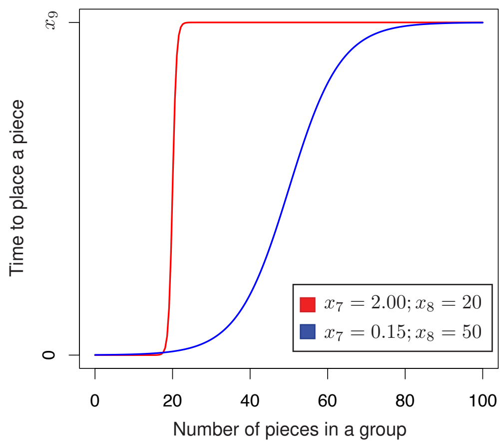

Figure 4: Time taken to place a piece

Generally speaking, the time taken to place all pieces in a particular group is simply the integral (from 0 to the total number of pieces sorted into the group) of the curve in Figure 4. Of course, placing pieces is harder if they have been incorrectly sorted. This is reflected in our model by reducing the mid-value of the growth phase of the logistic curve. For every incorrectly sorted piece, this value is reduced by 1 (until a minimum of 1 is reached).

We now have all the parts we need for our model. For completeness, we re-iterate that many of the model's assumptions could be questioned. In addition, it is difficult to say how many of the model's parameters could be affected by training. Despite those limitations, it remains an interesting model to analyse.

Investigation

For ease of reference, the parameters used in our model are shown in Table 1. This table also shows the range of values each parameter may take.

Table 1: Parameters used in the jigsaw model

	Interpretation	Min	Max
x_1	Number of groups	5	40
x_2	Power factor for group sizes	0.25	4
x_3	Growth rate for sorting time	0.25	4
x_4	Max. time to sort a piece	0.1	0.5
x_5	Growth rate for sorting probability	0.25	4
x_6	Min. correct sorting probability	0.8	0.95
x_7	Growth rate for logistic function	0.15	2
x_8	Midpoint for logistic function	20	50
x_9	Max. value for logistic function	5	10

The question we are concerned with is, which of these parameters will have the greatest effect on the time it takes to complete a jigsaw puzzle? There are various ways of investigating this question. For the purposes of this article, we use the Treed Gaussian Process (TGP) package for 'R: A Language and Environment for Statistical Computing'. As its name suggests, this package uses Gaussian processes, arranged in a tree structure, to create a fast-running emulation of a model; that emulation can then be used for various purposes, including sequential experimental design and sensitivity analysis. In our case, our model is so simple that we do not really need a fast-running approximation, we adopt that approach because it simplifies the use of TGP, which, in turn, facilitates the sensitivity analysis.

TGP allows two different types of sensitivity index to be calculated. There are *first-order* indices, which look at the effect of a parameter in isolation, and *total-effect* indices, which include interactions between one parameter and the others in the model. In our case, the first-order and total-effect indices show similar patterns so here we only report the first-order indices, which are illustrated in Figure 5. This figure shows a distribution of estimated first-order sensitivity indices in the form of a box and whiskers plot. Here, the bold line represents the median value; the rectangle extends from the first quartile to the third quartile; the lines cover a rough 95% confidence interval; and outliers are plotted as circles.

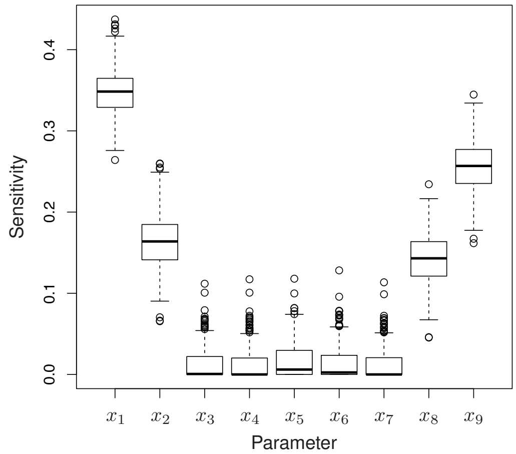

Figure 5: First-order sensitivity indices for the parameters in the jigsaw model

From Figure 5 it is apparent that the most significant parameters, in order of importance, are:

- x_1 – Number of groups
- x_9 – Maximum value for logistic function
- x_2 – Power factor for group sizes
- x_8 – Midpoint for logistic function

For completeness, Figure 6 shows the main effects for these four parameters. These plots show, in general terms, the effect of changing each of these four parameters. Note that the vertical scales differ between the plots; larger scales correspond to parameters that have larger sensitivity indices. Also note that the horizontal scales run from 0 to 1; this is a scaled version of the range shown in Table 1. Finally, note that the output of our model is the total time taken to place all of the pieces in the puzzle. Equivalently, lower values are better.

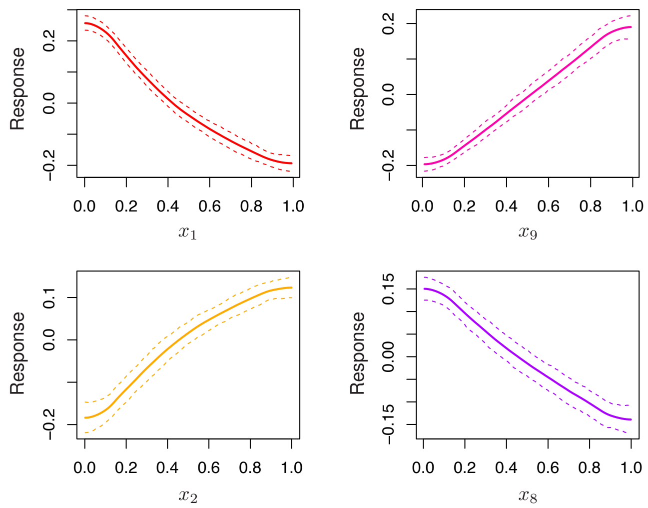

Figure 6: Main effects for the parameters in the jigsaw model

So, working across the individual plots in Figure 6, from left to right then top to bottom, if we want to complete a puzzle quickly we should: have as many groups as possible (x_1); make the maximum value of the logistic function as small as possible (x_9); make the scaling factor for group sizes as small as possible (x_2); and make the midpoint for the logistic function as large as possible (x_8). Intuitively, this makes sense because the aims related to x_1 and x_2 lead to small-sized groups, which are quicker to place, whilst the aims related to x_8 and x_9 constrain the effect of large-sized groups.

Revelation

So, what effect could a training regime based on these principles have? To briefly investigate this, we take as our benchmark the time taken to complete a 1000-piece puzzle when all parameters are at the midpoint of the ranges shown in Table 1. Against that baseline, the effects of sequentially applying the parameter changes identified above are shown in Table 2.

This table shows that, together, the parameters can have a significant effect, reducing the time taken to complete a puzzle to less than 30% of the time in the base case. It is interesting to note, however, that the biggest improvement comes from x_2, despite this having a lower sensitivity index than either x_1 or x_9. This apparent discrepancy arises because the calculation of sensitivity indices considers the full spread of possible values for the set of parameters, whereas each step in Table 2 is measuring sensitivity from a single point. This indicates the complexity that can hide even within a simple model. Practically speaking, it means that the benefit of a training regime depends on our current performance, or where we are starting from.

Table 2: Effect of parameter changes on time taken to complete puzzle

Case	Time taken
All parameters at midpoint	1.000
Then, x_1 maximised	0.803
Then, x_9 minimised	0.607
Then, x_2 minimised	0.299
Then, x_8 maximised	0.294

(In-)Completion

Although it has allowed us to gain some insight, this analysis is not likely to make me competitive at the British Jigsaw Championship. As well as my innate lack of skill, there are a number of potential deficiencies in the model: for example, we have no empirical evidence that demonstrates the model's validity; collecting such evidence (and refining the model, as appropriate) is left as an exercise for the interested reader. Likewise, the model does not account for pieces that are incorrectly placed within the puzzle.

Despite these limitations (and many others), I suspect that this article may have allowed me to achieve an ambition, albeit one that I have not held for very long. In particular, I suspect that this article makes me Britain's first 'jigsaw sports scientist', even if I have some way to go before producing anything of value to an elite jigsaw-ist!

Notes

More information on this year's British Jigsaw Championship is available from `www.achurchnearyou.com/church/2185/page/41400/view`. This year's Newmarket Jigsaw festival runs from 22 to 27 June, with the British Championship on 23 June.

More information on the R Project for Statistical Computing can be obtained from: `www.r-project.org`.

More information on TGP can be obtained from the relevant page of the Comprehensive R Archive Network (CRAN), specifically: `cran.r-project.org/web/packages/tgp/index.html`.

There are a number of alternatives to TGP that could have been used in this article. Of note is Dakota, which is produced by Sandia National Laboratories: `dakota.sandia.gov`.

Acknowledgement

'Urban Maths' cartoonist: Adrian Metcalfe – `www.thisisfruittree.com`

The views and opinions expressed herein are those of the author and do not necessarily reflect those of Dstl.