Modelling human problem solving with data from an online game

Tim Rach, Alexandra Kirsch

Cognitive Processing manuscript No.
(will be inserted by the editor)

Modeling Human Problem Solving with Data from an Online Game

Tim Rach · Alexandra Kirsch

Received: date / Accepted: date

Abstract Since the beginning of cognitive science, researchers have tried to understand human strategies in order to develop efficient and adequate computational methods. In the domain of problem solving, the Traveling Salesperson Problem has been used for the investigation and modeling of human solutions. We propose to extend this effort with an online game, in which instances of the Traveling Salesperson Problem have to be solved in the context of a game experience. We report on our effort to design and run such a game, present the data contained in the resulting openly available dataset, and provide an outlook on the use of games in general for cognitive science research. In addition, we present three geometrical models mapping the starting point preferences in the problems presented in the game as the result of an evaluation of the dataset.

Keywords Traveling salesperson problem · Problem solving · Casual Games

1 Introduction

One aspect of cognitive science research is the transfer of intelligent strategies from natural cognitive systems to computational systems. A fundamental challenge of this is to find adequate problems that are solvable and representable for both natural and computational systems, and to acquire enough data from natural systems to allow for identification of strategies and a basis for comparison.

T. Rach
University of Tübingen
E-mail: tim.rach@uni-tuebingen.de

A. Kirsch
University of Tübingen
E-mail: alexandra.kirsch@uni-tuebingen.de

In the area of problem solving, the Traveling Salesperson Problem (TSP) is a good candidate for such research, having been examined in psychology [14], as well as in computer science [5]. A TSP instance consists of a set of points, where the task is to find the shortest tour that visits each point once and returns to the origin. Though being an NP-hard problem, humans generally produce remarkably good solutions [14] [22], which makes it an interesting object of study on problem solving. However, experimentation is time consuming and cost intensive: participants have to be found and supervised, materials printed and analyzed.

To avoid these steps as well as an artificial laboratory situation we developed an online game, mapping planar Euclidean 2D instances of the TSP to a playful task in an appealing surrounding. We chose to use an online game because researchers have successfully used games in the past to acquire data of human behavior in different tasks [20, 2, 8] with compelling results.

The goal of this paper is to share our experience with designing and running an online game to gather data on human problem solving behavior[1] and to use that data to improve models simulating human solution strategies in the domain of the TSP. There are different approaches for such simulating models, covering basic strategies (Cutini [3]), hierarchical approaches (Kong & Schunn[9], Best[1], MacGregor[16], [18]) and combinations of those (Kirsch [7]). In some models, the set of possible starting points is implicitly defined by the definition of the represented strategy, while in other models the set of possible starting points includes all points of

[1] The game can be played at `www.perlentaucher.medieninformatik.uni-tuebingen.de` and the resulting dataset is available at `http://www.wsi.uni-tuebingen.de/lehrstuehle/human-computer-interaction/home/code-datasets/tsp-dataset.html`.

the problem. However, none of the existing models use a strategy based on test results to choose a specific point as the starting point for the constructed tour, but a random point in the set of possible starting points. For some of the models this results in different tours when applying the model multiple times to the same problem. MacGregor[12] observed, that points on the convex hull and points near the geometrical center or the center of mass of the problem are often chosen as starting points by humans. Also, starting point preferences varied across individuals, ranging from 7% to 100% for the frequency of hull starts. Those results indicate, that in general the convex hull might be a good selection as the set of possible starts for a model simulating human behaviour in TSPs, but disregard the cases where interiour points are preferred. Other models ([9] , [1]) choose a cluster based approach for the selection of the tour start, which fails for small instances of the TSP, as the presence of visible clusters is not ensured. Although there is no study that clearifies the relevance of starting points for the cognitive process of tour production, previous unpublished data showed strong preferences for specific points in TSPs. In the analysis in section 4 we use the large amount of data gathered in the game to identify possible factors for those preferences and introduce three geometrical models that could be used to predict preferred starting points. For a literature review on human performance in solving TSPs and suggested models, we recommend the review by MacGegor and Chu[14].

2 Game Design

Our main goals in developing this game were:

1. to address a large number of participants and create an appealing game experience that does not feel like a test,
2. to extend the available observations on TSP solving with a wider range of options for the solution process and different problem variants.

To achieve the first goal, we wanted to create a casual game. Basic principles of casual games are “easy to learn, simple to play and offer[ing] quick rewards with forgiving gameplay” [11], with popular examples like *Angry Birds*©, *Cut the rope*© or *Candy Crush Saga*©.

Our game *Perlentaucher*[2] (Pearl Diver) introduces the TSP in a simple story: the pearl-diving panda Paul wants to make the process of collecting pearls more efficient and therefore needs to find the shortest tour on each of his diving spots. The story is presented in an introductory level that also explains the game mechanics, the user interface and the functions of purchasable game advantages (see Sections 2.1 and 2.2).

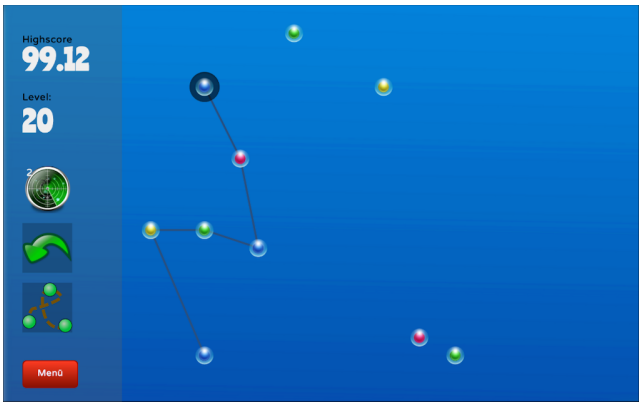


Fig. 1 A level with differently colored nodes

The setup of a casual game is well suited to achieve the second goal: as is usual in games, our players receive feedback and can repeat levels. This provides data on learning or optimization effects when solving a problem instance several times. This aspect is very important for computational models for human TSP strategies, because all models we are aware of (e.g. The Sequential Convex-hull Model, Pyramid Models, Global-local Models [14]) produce exactly one solution. Another dimension is to examine the influence of tools or hints to solve TSP instances. *Perlentaucher* includes three purchasable game advantages with possible aids, offering not only the possibility to observe whether such aids help to find the optimal solution, but also how useful the players estimate the aids to be (as they have to be purchased with points). A third dimension for variation are the levels themselves. Of course, each TSP instance is different, but we also vary the task slighly in different blocks of levels (see Section 2.3).

2.1 Board

The levels are displayed on a blue background representing the ocean (Figure 1) with the nodes shown as images of pearls, which the player has to collect on the shortest path. The pearls are drawn onto a grid of 20 rows and 26 columns to allow for a better distinction of the distances between the pearls. Already collected pearls are connected by lines representing the chosen tour with the last collected pearl marked by a dark circle. To avoid confusion and illegal tours, every pearl except the starting point can be selected only once. In addition, the tour can be closed only when all pearls are included.

[2] The game is in German as to give us better access to local players. An english version is in development.

Fig. 2 The level representation within the areas

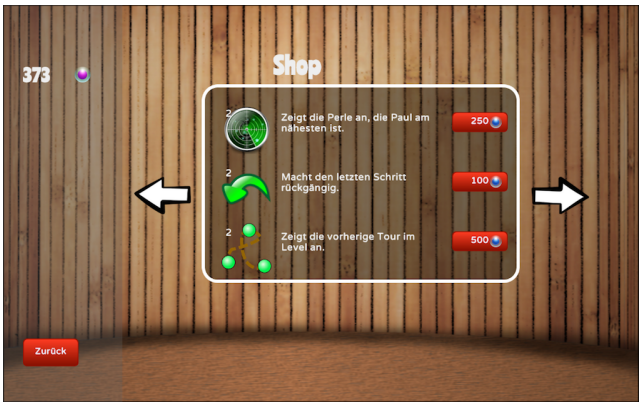


Fig. 3 The game's shop area

Fig. 4 The area-selection screen

The sidebar on the left shows the player's best score from previous runs, the level number and the available game advantages.

When the player finishes a level, a pop up dialogue shows the rating, which is caluclated as the ratio of the shortest possible tour length and the found tour length scaled by 100, resulting in 100 points when the optimum was found and accordingly lower scores for longer tours. Corresponding to the score, the player is rewarded with a bronze (score ≥ 90), silver (score ≥ 95) or gold coin (score $= 100$). Earned coins are shown in the level selection screen to provide an overview of the player's progress (Figure 2). To be able to play a level, the player has to finish all previous levels with a score of at least 90 points.

2.2 Game Advantages

The player may use three items in the game that possibly help to find the optimal solution:

- Undo: reverts the last node selection
- Nearest neighbor: shows the nearest node to the currently selected node
- Show last tour: shows the last complete solution of the user for the current level.

The functionalities of those items are the result of a user test where participants were asked which functions they considered helpful for finding the optimal tour. The items have to be purchased in the game's shop area (Figure 3) with points the player has received by solving levels. Some items are provided for free intially to allow the players to use the items from the very start.

2.3 Level Construction

The game contains 24 levels which are devided into three groups of 8 levels, each representing different variations of the TSP. The groups are represented by areas which have to be unlocked by finishing the previous area (Figure 4).

The first group contains plain Euclidean TSPs: Free choice of a starting point with all nodes (pearls) having the same color.

In the problems of the second group again all nodes have the same color, but the starting point is preselected and cannot be altered by the user. In some levels the starting point is one that has been often chosen in previous experiments, in others one that has only rarely been chosen.

In the third area the nodes are marked with different colors. The users are informed of the nodes being colored, but recieve no further instruction. For each level, three to five colors are used and distributed over the nodes in three ways:

- colors emphasize visible clusters, each cluster being marked with a specific color;
- same colors are used for neighbouring nodes, defining regions, but not following visible clusters;
- colors are randomly assigned to nodes.

For the first two coloring schemes there are two kinds of levels to see whether the coloring influences the human solution strategy:

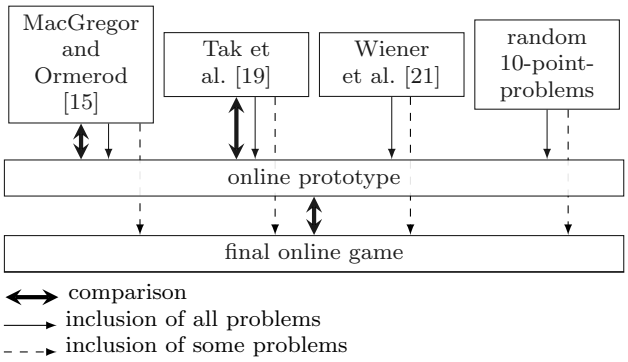


Fig. 5 Overview of used TSP instances and the comparison of results

- following, i.e. visiting all nodes of the same color before moving on to another color, helps the user to find the optimal solution;
- following regions misleads from the optimal solution.

Some TSP instances are used in two different levels, allowing some direct comparison of conditions.

2.4 Data

The following data are recorded when a player solves a level:

- timestamp of when the solution was saved to the database
- tour as a list of node numbers
- tour length
- time in seconds from starting to completing the tour
- level number
- user-ID
- used game advantages

To limit concerns about privacy and the necessity of additional security techniques, we do not record any information about the user's input- or display devices as well as the user's age or gender. This may result in some variation of the data, especially for the duration to complete a tour. It should be noted that the results for the solution times should be treated with caution, as this component is highly dependent on the used input device. Given that, the tracked times represent more the time needed to submit the solution than the time needed to actually solve the problem. Nevertheless we included the time in the available data sets, as they might be useful for some evaluations.

3 Observations

With the online game we sacrifice the controllability of a lab experiment. This is why we wanted to test the reliability of this form of data acquisition by comparing it to previous results from laboratory experiments.

The game was developed in a user-centered design process. The last prototype contained — among others — TSP instances from experiments by MacGregor and Ormerod [15] and Tak et al. [19]. The prototype game was played online by 27 participants (mostly students of computer science) over 10 days. Most of the levels used in this user test are no longer present in the currently used level set as they did not follow our construction directives. This is why we compared the results obtained from the last prototype with those reported in the literature and in a second step looked at the problems from this prototype that are still used in the final version (see Figure 5).

A standard measure for comparing TSP solutions is the percentage above the optimal tour length (PAO). The comparison of the 18 problems by MacGregor and Ormerod [15] and Tak et al. [19] is shown in the Appendix B, Figure 13. Because the participants in those studies could solve each TSP instance once, we only use the first trial of the players for each level from our game data. The PAO values of our data differ at most by 2.78 from the reference data of MacGregor [15] and at most by 2.4 from the reference data of Tak [19]. On average the differences are 1.34 (MacGregor) and 1.08 (Tak).

Figure 6 shows the comparison of the PAO values of the first solution attempts in the last prototype version and in the current dataset for the seven problems shared by both versions. The problems are specified in appendix, Table 3.

In both steps, the results are comparable, which shows that the noise in our online game is at an acceptable level. Although the maximum PAO values in the current data are much higher than in the prototype, the results are similar, as the large size of the data set compensates for those outliers.

In addition, we wanted to know if we achieved the original goals of our project.

Goal 1: to address a large number of participants, creating an appealing game experience that does not feel like a test. During the development phase, we got positive feedback from our participants on the game design and its entertainment value. We have gathered new data from $38,400$ games, played by over $1,200$ different players. The participants were recruited by email advertising to students and employees at the University of Tübingen and to personal acquaintances, which lead to approximately $30,000$ solutions in a few days. The game is still played regularly, but not as often as when we started to make the game known. With a real ad-

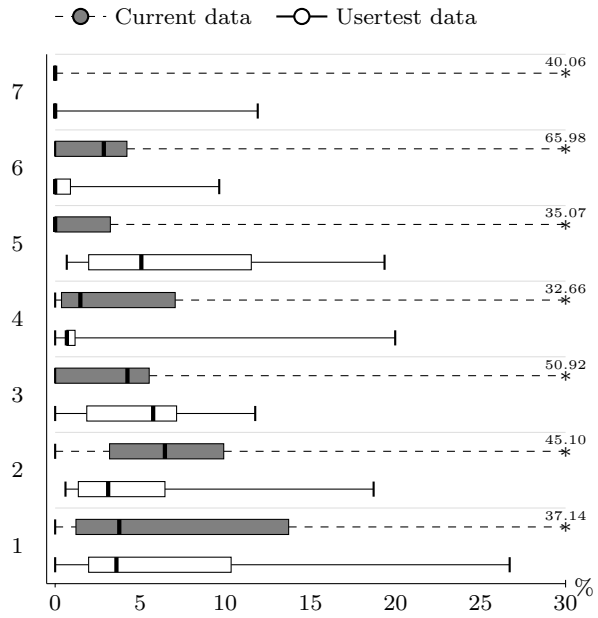


Fig. 6 Comparison of PAO values of 7 problems used in the prototype and in the final version. The whiskers show the minimum and maximum values, the box contains values between the upper and the lower quartile, the bar indicates the median value. The game data contains outliers with PAO values beyond the scale of this figure, indicated with a star and the maximum value.

vertising campaign, the range of players could probably be enlarged.

Players can also give feedback in the online game. Most comments are positive with respect to the entertainment aspect. Suggestions for improvement include a mobile version for smartphones, additional game advantages and more challenging levels.

There seem to be different groups of players: some are just curious and explore the levels once or twice without trying to find the optimal tour, while other players are really ambitious to solve the whole game (from personal feedback we know that some players even play against time when they have already found the optimal solution). From their perseverance and positive feedback we conclude that at least the second group really enjoys the game. 33 players found the optimal solutions for all 24 levels.

Goal 2: to extend the available observations on TSP solving with a wider range of options for the solution process and different problem variants. By construction, the game offers many options for variation as described in Section 2. We have just started to analyze the data, but we already made some interesting observations:

- Predefined starting points seem not to impair the solution quality, even when participants are forced to start with a point that they would never choose voluntarily.
- Getting feedback and having several attempts to solve the problem leads to significantly better solutions than the first one. In this light, humans are even better in solving (small) TSPs than has already been known from the first attempts. But, in the course of finding the optimum, participants can also produce solutions that are worse than the first attempts.
- The game advantages are used rarely, even though they were suggested by our test users in a previous user study. It remains to be analyzed how much such hints help to find the optimal tour.

4 Analysis of Starting Point Selection

4.1 Participants and Stimuli

Participants were mainly students and employees at the University of Tübingen that were recruited via email (cf. Section 3). The participation was voluntary and not rewarded. As the registration and participation was freely accessible from the internet and we invited the players to pass on the link to their acquaintances, it is likely that participants of other groups than the mentioned have played the game.

The presented stimuli were 24 instances of the Euclidean TSP varying in size from 5 to 20 points, each belonging to one of the four categories described in section 2.3. It was not required to solve all problems, which resulted in varying numbers of participants among the problems. The maximum number of participants was 1238 at the first problem and the minimum number 629 at problem 24.

4.2 Procedure

The game was playable from devices running a modern browser and required a mouse or touchpad as input device, touchscreen devices were not supported. The participation required the registration and the completion of a tutorial explaining the task. The presented instances had to be solved in a fixed sequence that was equal for each participant and allowed to solve a problem multiple times and to go back to previous problems.

Statistical method As the purpose of statistical evaluation was to identify selection frequencies that differed significantly from chance, the statistical method of choice was the binomial distribution. The quality of

starting points was determined by converting the start counts of each point to standard scores for the normal distribution approximating the underlying binomial distribution. A detailed explanation is given in the 'Calculating Standard Scores with the Binomial Distribution' sidebar.

4.3 Analysis Results

The experiment resulted in a total of 38465 tours, of which 20770 were first attempts. For the comparability with other studies where in general the problems are solved only once by each participant, we only took the first attempts (8936 tours) of the first 8 problems into account for the evaluation of significant preferences. The remaining 16 problems were not taken used for the analysis of features, as they had preselected starting points or differently colored points, which could affect the choice of the starting point.

Presence of preferred points: At least one highly significant point was found in each problem ranging from 12.4 σ to 56.1 σ. In total 20 of the 96 possible points qualified as a possible starting point by being beyond 5σ. (We chose $+5\sigma$ as the threshold for significant starting points, as values above that threshold stand for the upper 0.0001% of the distribution, which we considered precise enough to rule out chance.) The results not only support the reports of preferences for points located on the convex hull and centeroid points, but also indicates, that in those regions preferences for single points exist.

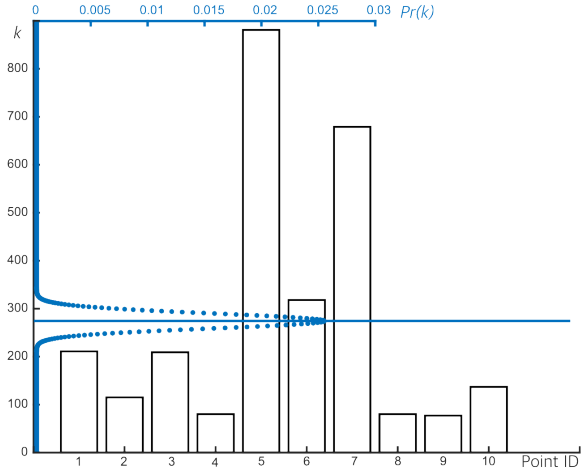
Reasons for preferences: MacGregor [13] proposes an initial contour, effectiveness and visual salience as reasons for humans preferring boundary points over interior as tourstarts. We adapted the thought and tested wether the effectiveness or saliency explanation applied for the points preffered by participants in each level. The initial contour explanation was not tested as it generalizes the properties of preffered points too much.

If starting at a specific point were more effective, the resulting tours should be shorter in length or the time needed for the solution process than tours starting at other points. To resolve that question, we calculated the PAO for every tour and the σ of the corresponding starting point. Figure 9 shows the mean PAO of tours starting some point in correlation to the σ value of that point. The calculated correlation coefficient $\rho = -0.04$ (Spearman's correlation coefficient) between the relative deviation of a point and the corresponding PAO

Calculating Standard Scores with the Binomial Distribution

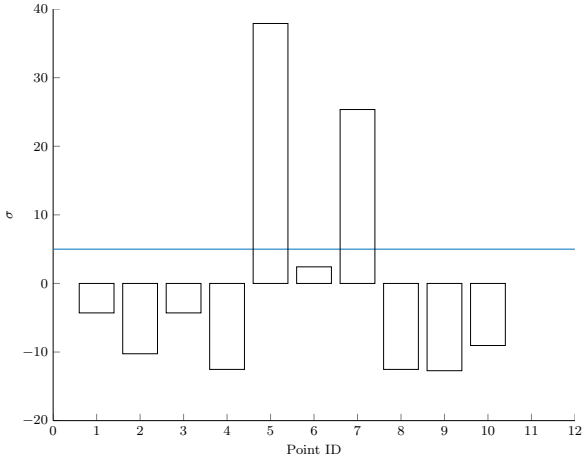
The binomial distribution describes the probability Pr of obtaining exactly k successes in a set of n independent experiments where the result is either success (with probability p) or failure (with probability $1 - p$).
An example is shown in Figure 7: It shows the results of the evaluation of tour starts in a TSP instance containing 10 points. As the participants were free in the choice of a starting point, each point has the probability $p = \frac{1}{10}$ of being selected as a start. 2768 tours were produced for the problem, which leads to a probability distribution with the parameters $n = 2768$ and $p = \frac{1}{10}$. The visual representation of the resulting probabilities is displayed in blue on the y-axis with the mean at $n \cdot p = 276$.

Fig. 7 Numbers of tourstarts at different points and the corresponding binomial distribution



If n is large enough, which in this case is given, the binomial distribution approximates a normal distribution. That allows the conversion of the number of starts at each point to the number of standard deviations above the mean (standard score).
The corresponding standard scores for the given example are displayed in Figure 8.

Fig. 8 Standard scores for each point in the given example. The blue line marks the 5σ threshold.



values of the tours starting at that point did not indicate any monotonic correlation between the two dimensions. The evaluation of solution times in correlation to the σ values of points also did not show any correlation ($\rho = -0.01$) - As mentioned before, the result for the solution time should be treated with caution, as some uncontrollable factors influence the captured solution time, e.g the used input device.

The low correlation values contradict the theory of specific points leading to better tours.

The second possible reason for preferences is, that specific points are visually more salient than others. Visually salient items grab the attention by differing in one or more attributes from their neighbors. In [23] 5 attributes are named that guide visual attention: color, motion, orientation and size. As the points were all presented as static circles in the same color and size, none of those attributes could be used to extract visually salient points.
For the further evaluation we used other attributes that we think may be sources of attraction of attention:

1. The position of the point in the problem: Of the 20 significant points, 17 were located on the convex hull. The other 3 points were all closest to the geometrical center of the problem and 2 also closest to the center of mass. In total 74% of the tours produced in the initial attempt and 72% of all tours were started on a boundary point, while the rate expected by chance was 61%.
   Only 4 problems contained a point in close proximity to the geometrical center, of which 3 were significant. The instances that did not contain a centeroid point showed a visible, though not significant, preference for points in the centeroid region.

2. Salient structures: Points that form prominent structures such as lines, triangles, circles, squares etc. in a problem, could be candidates for starting points as the structure itself may draw visual attention. See Figure 10 for an example: The problem contains a straight line formed by 7 points; As the structure sticks out, the points in it could have a higher chance to be picked as a starting point than the points not included in the line.
3. Spatially isolated points: Points with few neighbors in close proximity appear to be isolated and therefore could be visually salient.

Fig. 9 Correlation between the σ-value of a point and the mean PAO of all first-attempt tours starting at that point

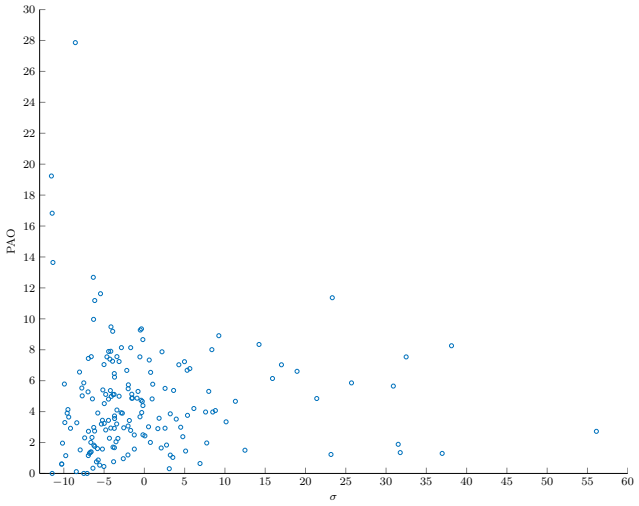
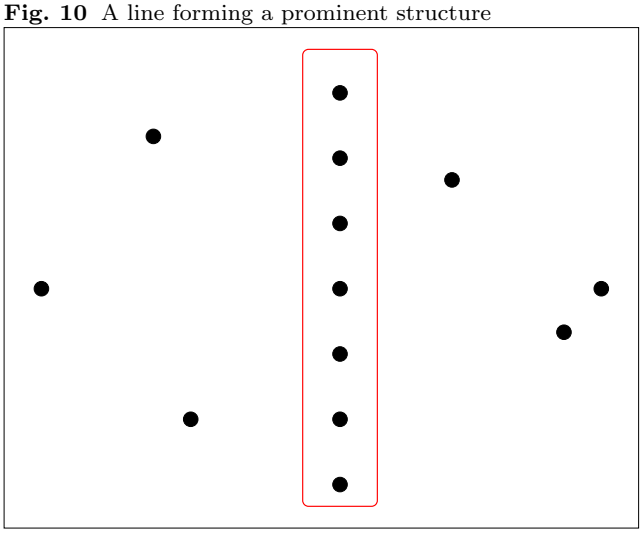


Fig. 10 A line forming a prominent structure

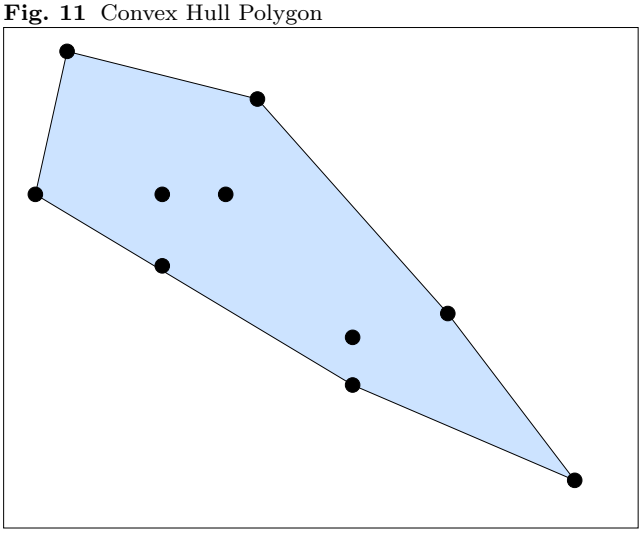


5 Models for start selection

Based on the observations described in section 4.3 we derived three models that can be used to identify points that may be favored over others as tour starts.

1) Smallest Angle (Algorithm 1) The smallest angle model is based on the assumption, that points gain salience by being part of a prominent structure. Since the human visual perception follows a top down process recognizing global features before fine-grained features [17], the first recognized shape when looking at a TSP is the polygon formed by the convex hull. As mentioned earlier, not all points of the set defining the convex hull were significantly often chosen as a starting point, which is why we tried to identify salient points within this set. For this model we chose the angle be-

Fig. 11 Convex Hull Polygon



tween the lines connecting a point with its neighbors as the saliency feature. We consider points where the inside angles of the polygon shape are relatively small as more salient than points where the angle is rather wide. Figure 11 shows an example: The angles at the topmost and the bottommost point of the problem are much smaller than the other angles in the polygon, which makes those two points more salient than the others.
In the experiment the identified points were chosen significantly more often than by chance: The topmost point scored $\sigma = 23.3$ and the bottommost point $\sigma = 32.4$, while all other points scored below zero, i.e were chosen less often than by chance.

Algorithm 1 Smallest Angle Model

```
1: procedure SMALLEST ANGLE(Points)
2:     K := convexHull(Points)
3:     for p ∈ K do
4:         v1 = p − neighbor_right
5:         v2 = p − neighbor_left
6:         θ := acos((v1 · v2)/(norm(v1) ∗ norm(v2)))
7:         if θ < θ_min then
8:             θ_min := θ
9:             start := p
10:        end if
11:    end for
12:    return start
13: end procedure
```

2) Maximum Distance (Algorithm 2) This very simple model tries to identify the most isolated point in the complete point set by accumulating the distances to all other points. The point with the greatest accumulated distance is considered the most isolated one and therefore is identified as a possible starting point. It is possible, that multiple points have the maximum accumulated distance. In that case, both equally qualify as a starting point.

Algorithm 2 Maximum Distance

```
1: procedure MAXIMUM DISTANCE(Points)
2:     distances := [0, . . . , 0]
3:     for i := 1 to |Points| do
4:         for j := 1 to |Points| do
5:             distance[i] += norm(Points[i] − Points[j])
6:         end for
7:     end for
8:     returnPoints[distances.indexOf(max(distances))]
9: end procedure
```

3) Relative Maximum Distance (Algorithm 3) The third model is similar to the previous one, but much more complex: While the Maximum Distance model finds the most isolated point for all n neighbors present in the problem, this model finds the most isolated point for each number of neighbors. To do so, for each point the radii $r_1 \dots r_{n-1}$ are calculated for a circle with radius r_i around the point to include i neighbors. By counting how often a point has the greatest distance to its $1 \dots n-1$ neighbors, the (in terms of this model) most isolated point can be determined. The point that has most often the smallest distance to its neighbors is the point that lies nearest to the center of the problem, which is also useful as those points were also chosen as starting points. For the evaluation in Table 1 we counted for each point how often it had the greatest distance (GD score) as well as how often in had the smallest distance (SD score) to its neighbors. The point with the highest score, regardless of GD or SD , was identified as the winning point. If two scores (GD and SD) were equal, the GD-score was favored, if two GD scores were equal, the winning point was chosen randomly.

5.1 Model performances

The models assign scores to the points, with the point having the highest score being regarded as the most probable starting point. This method seems to allow for a ranking of the points, i.e. the point having the second highest score also being the second most probable starting point. Unfortunately the data (excluding the highest scoring point) does not confirm a linear correlation between the ranking of a point and the actual start frequency.

Algorithm 3 Relative Maximum Distance

```
1: procedure RELATIVE MAXIMUM DISTANCE(Points)
2:     for i := 1 to |Points| do
3:         radii(i) := calcRadii(Points[i], Points)
4:     end for
5:     counts := [0, ..., 0]
6:     for d := 1 to |Points| − 1 do
7:         distances := []
8:         for r := 1 to |radii| do
9:             distances[r] := radii[r][d]
10:        end for
11: Find the Points with the greatest/smallest distance to
    their d next neighbors.
12:        farthest := max(distances)
13:        nearest := max(distances)
14: Increment the corresponding count values - If multiple
    points apply for farthest or nearest, increment all of them.
15:        counts[distances.indexOf(farthest)] += 1
16:        counts[distances.indexOf(nearest)] += 1
17:    end for
18:    start := Points[counts.indexOf(max(counts))]
19:    return start
20: end procedure
21:
22: procedure CALCRADII(p, Points)
23:    for i := 1 to |Points| − 1 do
24:        radii(i) := distance to i-th nearest neighbor of p
25:    end for
26:    return radii
27: end procedure
```

Table 1 shows the prediction results of the models for the evaluated eight instances: a ranking of 1 means that the model predicted the point with the highest standard score - the point that was selected most often as a starting point. Accordingly stands a rank of 5 for the point with the fifth highest standard score. Rankings marked with the $\sim$ stand for points where the rate of starting selections did not significantly differ from chance.
The smallest angle model relies completely on the differences between the angles at the points on the convex hull, which makes it rather useless for problems where all angles are similar, e.g. problems where the hull forms a circle, a rectangle or a triangle (Levels 5 and 7).
Another problem that is shared by the Maximum Distance model is that interior points, especially the point nearest to the center, are not considered as possible starts which results in accordingly bad predictions (Level 7). Too few points in a problem lead to the failure of all models - the problem (level 2) is shown in Figure 12

In the model desciptions above, only one point - the point that fits the models criteria best - is returned as the result, which does not apply in reality: our data contained problems with up to four points with a σ value greater than 5. A solution for that can be to return a number of points with the highest scores within the model.

Fig. 12 The second level in the game. The relevant geometrical properties are almost equal for all points, which is why none of the models could identify the point chosen most often as a starting point by the participants (Point 1).

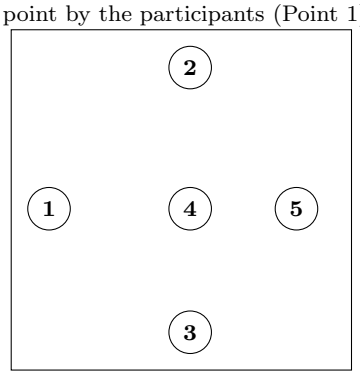


Table 1 Rankings of the models. Entries marked with $\sim$ stand for predictions with $\sigma \leq 5 \rightarrow$ bad predictions

Level	Smallest Angle	Maximum Distance	Relative Maximum Distance
1	2	2	2
2	5˜	5˜	3˜
3	1	1	1
4	1	1	1
5	2	1	1
6	1	1	1
7	4	6˜	1
8	1	1	1

6 Other observations

Perceptional salience: The salience criteria used in the analysis are not listed as 'undoubted attributes'[23] for visual attention and therefore cannot be used as strong arguments for the saliency explanation. In the set of TSP instances one problem was used twice in different versions: The first being a plain problem, the second forming groups of nodes by applying a different color to each group. A noticable difference in the start selection behavior was the shift of start selections to a point having only differently colored points as neighbors. Although the shift is only small, it might indicate that starting point selections can be influenced by saliency features such as coloring.

Relevance of starting points: In section 4 we mentioned that the relevance of the starting point for tour production is not clear. Ten TSP instances in our test set had preselected starting points. Two of those instances were also included as a version without a preselection. Table 2 shows the comparison of PAO values for those instances. The comparison of level 3 to level 12 shows better results for the version with the fixed starting point, which could mean that the preselection had a

positive effect on the performance (regarding the minumum and mean values) but it is more likely that some participants recognized the level and the relatively easy solution (comments in the feedback support this theory). The comparison of the levels 7 and 14, which have a rather difficult solution, shows almost no differences except for the maximum value. Assuming, the PAO is a good method for the comparison of tour qualities, the results deny a relevant role of the starting point in the construction process of tours.

7 Discussion

The *Perlentaucher* game is an attempt to facilitate the acquisition of data about human problem solving. It is not the first time a game is used to collect data, but it is rather unique in its design as a casual game. In a similar line, well-known games such as Tetris [6] or Angry Birds [10] have been suggested to study human cognitive abilities. The advantage of *Perlentaucher* is that we can easily access the data, change the game and do not have to deal with any copyright issues. But, the game as such first has to be explained and may not be as interesting as well-known popular games.

A potential drawback of data gathering with online games (no matter if the game is well-established or newly introduced) is the low level of controllability, resulting in data outliers. As mentioned before, different hardware can cause different performance in the game, especially for solution times, but also the contrast or brightness of a monitor may result in a different perception of the task and lead to distorted results.

Besides the used hardware, the environment and distraction level of the players are completely unknown and can hardly be controlled. As we have seen in *Perlentaucher*, the motivation level of players can also vary significantly. In a game that is not previously known to players, one can also not completely ensure that the task is well-understood. The feedback data did include one comment (of over 600) where the user did not understand the task at first. We tried to compensate for this with the obligatory instructional level but even in laboratory tests one cannot guarantee that the instructions are fully understood, even though the participants would have more options for clarification. In this case the user did resolve the problem on his/her own by absolving the instructional level a second time.

The variance induced in the data by all these factors can to some extent be compensated by the larger amount of data that can be acquired with an online game. Our tests in *Perlentaucher* showed comparable results to previous studies from the literature, but we cannot generalize this to other games.

Along this line, one has to be aware that designs from laboratory studies cannot be taken over directly to online games. We have a between-subjects design, where all subjects solve the same problems, but we cannot use counter balancing for the order in which the tasks are presented. It is possible to randomize the order of the levels, but it may feel odd to the players and in the case of *Perlentaucher* destroy the grouping of levels into diving areas. Also, the possibility of directly testing the influence of a specific independent variable is limited. We have used some TSP instances for two levels to make some comparison possible, which was noted by seven players in the open feedback section (it may have been noted by other players as well). This repetition was mentioned rather as confusing than annoying; in any case such repetitions should be used with care. Too many and too obvious repetitions definitely diminish the entertainment factor of any game. Another option is to program the web server to provide slightly different conditions to different players (for example, in *Perlentaucher* different players could see the same problem, but in different orientations). One would have to ensure that the players always see the same version of the level (which would be possible), but players may still feel annoyed when they accidentally find out that other players got the same level in a different version.

We compared the results obtained with *Perlentaucher* to small-sized problems from the literature. In some studies [4], TSP instances of more than 100 points were used. Given the design of the game and available space on the screen, we cannot test such large instances. However, the TSP instances humans solve in everyday life (such as shopping tours or vacations) are rather in the order of the problems included in *Perlentaucher* and are thus interesting objects of study.

The large number of solutions allowed for a more thorough analysis of chosen starting points and the effect of starting points on the solution. We compare several models to extract salient points from the problems and depending on the structure of the problem (circular, point near the center, etc.). The comparison of the models to chosen starting points confirms that saliency plays an important role in the choice of the starting point. However, the performance of each individual model depends on the specific TSP instance. This issued might be solved with some classification of problem structures (similar to [3]), but as the choice of the starting point is subject to individual preferences, a set of models may be a more realistic explanation.

Despite the simplicity of the used geometrical features, the good predictions for the provided problems could be the result of overfitting. We could not test the models with other TSP instances, because the data

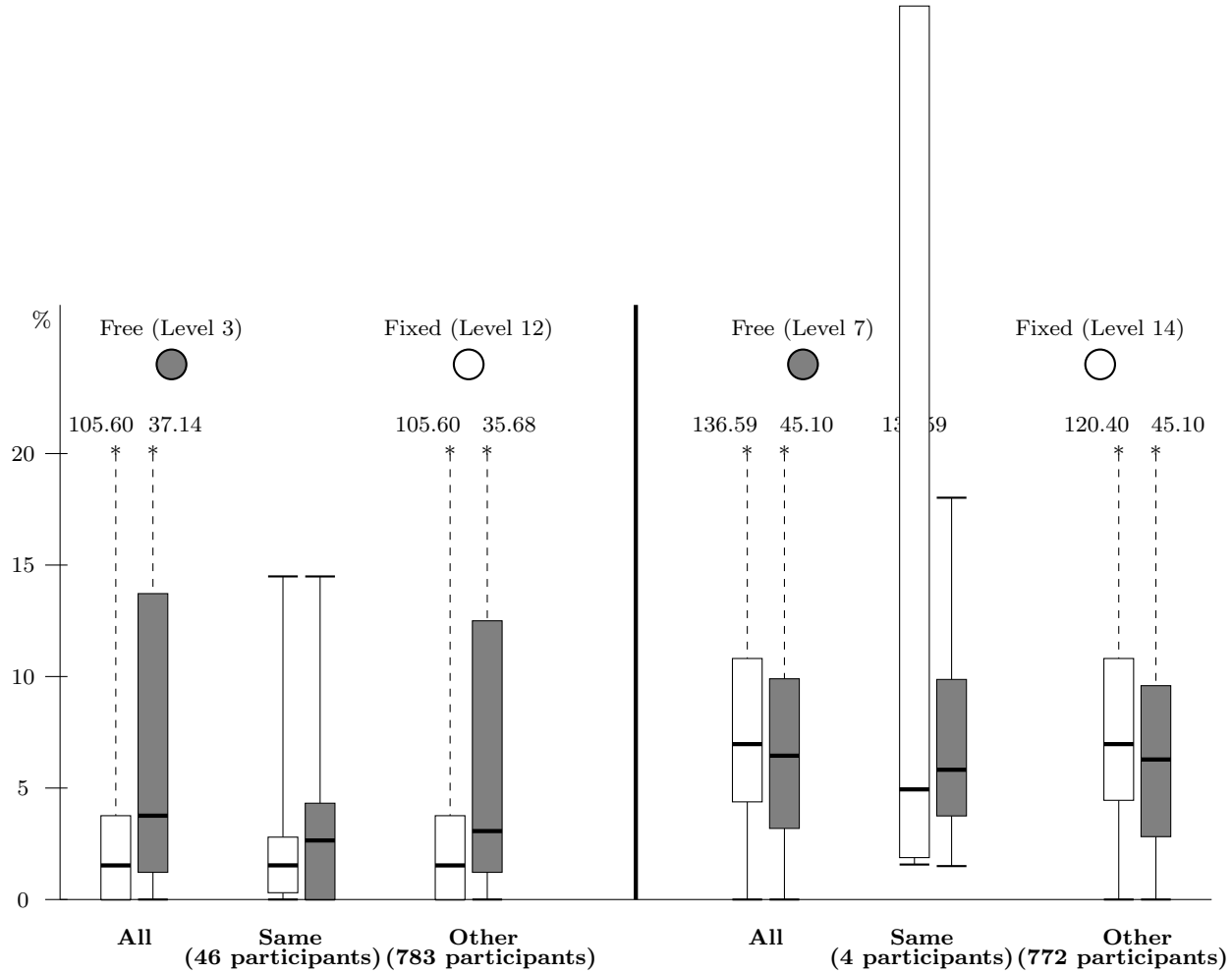


Fig. 13 Comparison of PAO values for geometrically identical problems in two versions: Free choice of tour start (Free) and with a preselected tour start (Fixed). Values are shown for participant groups **Same:** Participants who chose the same point for their tour start in the free version as the one that was preselected in the fixed version of the problem; **Other:** Participants that chose any other point as tour start in the free version; **All:** All participants of the two other groups combined.

we found either did not record the starting point of the provided solutions or the size of the dataset was too small to show significant selection rates. We plan to add more levels to the *Perlentaucher* game and use those as independent test instances.

Our data also suggests that the starting point is more likely the result of saliency properties than the result of playing an important role in the cognitive process of tour production. Nevertheless, the data showed significant preferences for specific points and therefore a selection strategy for starting points in models simulating human solution behaviour for the TSP is reasonable.

8 Conclusion

From our experience with *Perlentaucher*, we can recommend online games as an experimental method. It offers large flexibility in the design and variation of the tasks and is an effective way to record large datasets quickly and with low organizational overhead.

Using this large dataset we could analyze the choice of the starting point in greater detail. But there are many more open questions such as the extent to which people learn from prior trials or interindividual differences when solving TSPs, which we an other researchers can tackle with this data.

We hope that the data we have been gathering with *Perlentaucher* will be useful to the community to further explore human problem solving and we appreciate feedback for improvement of the game and further TSP variants to include as new levels.

Acknowledgements

With the support of the Bavarian Academy of Sciences and Humanities.

References

1. Best, B.J.: A model of fast human performance on a computationally hard problem. In: Proceedings of the 27th annual conference of the cognitive science society, pp. 256–262 (2005)
2. Brozik, D., Zapalska, A.: The restaurant game. Simulation & Gaming **31**(3), 407–416 (2000)
3. Cutini, S., Di Ferdinando, A., Basso, D., Bisiacchi, P., Zorzi, M.: A computational model of human planning in the traveling salesman problem. In: Proceeding the 27th Annual Cognitive Science Conference, pp. 524–529 (2005)
4. Dry, M., Lee, M.D., Vickers, D., Hughes, P.: Human performance on visually presented traveling salesperson problems with varying numbers of nodes. The Journal of Problem Solving **1**(1), 4 (2006)
5. Golden, B., Bodin, L., Doyle, T., Stewart, W.: Approximate travelling salesman algorithms. Operations Research **28**, 694–711 (1980)
6. Gray, W.D., Perez, R., Lindstedt, J.K., Skinner, A., Johnson, R.R., Mayer, R.E., Adams, D., , Atkinson, R.: Tetris as research paradigm: An approach to studying complex cognitive skills. In: P. Bello, M. Guarini, M. McShane, B. Scassellati (eds.) Proceedings of the 36th Annual Conference of the Cognitive Science Society. Cognitive Science Society, Austin, TX (2014)
7. Kirsch, A.: Humanlike problem solving in the context of the traveling salesperson problem. In: AAAI Fall Symposium: Advances in Cognitive Systems (2011)
8. Kirsh, D., Maglio, P.: On distinguishing epistemic from pragmatic action. Cognitive science **18**(4), 513–549 (1994)
9. Kong, X., Schunn, C.D.: Global vs. local information processing in visual/spatial problem solving: The case of traveling salesman problem. Cognitive Systems Research **8**(3), 192–207 (2007). URL `http://d-scholarship.pitt.edu/22731/`
10. Kreutzmann, A., Wolter, D.: Physical puzzles—challenging spatio-temporal configuration problems. In: Proceedings of IJCAI-2011 Workshop Benchmarks and Applications of Spatial Reasoning (2011)
11. Kuittinen, J., Kultima, A., Niemelä, J., Paavilainen, J.: Casual games discussion. In: Proceedings of the 2007 conference on Future Play, pp. 105–112. ACM (2007)
12. MacGregor, J.N.: Indentations and starting points in traveling sales tour problems: implications for theory. The Journal of Problem Solving **5**(1), 3 (2012)
13. MacGregor, J.N.: An investigation of starting point preferences in human performance on traveling salesman problems. The Journal of Problem Solving **7**(1), 10 (2014)
14. MacGregor, J.N., Chu, Y.: Human performance on the traveling salesman and related problems: A review. The Journal of Problem Solving **3**(2) (2011)
15. MacGregor, J.N., Ormerod, T.: Human performance on the traveling salesman problem. Perception & Psychophysics **58**(4), 527–539 (1996)
16. MacGregor, J.N., Ormerod, T.C., Chronicle, E.: A model of human performance on the traveling salesperson problem. Memory & Cognition **28**(7), 1183–1190 (2000)
17. Navon, D.: Forest before trees: The precedence of global features in visual perception. Cognitive psychology **9**(3), 353–383 (1977)
18. Pizlo, Z., Stefanov, E., Saalweachter, J., Li, Z., Haxhimusa, Y., Kropatsch, W.G.: Traveling salesman problem: A foveating pyramid model. The Journal of Problem Solving **1**(1), 8 (2006)
19. Tak, S., Plaisier, M., van Rooij, I.: Some tours are more equal than others: The convex-hull model revisited with lessons for testing models of the traveling salesperson problem. The Journal of Problem Solving **2**(1), 2 (2008)
20. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 319–326. ACM (2004)
21. Wiener, J., Ehbauer, N., Mallot, H.: Planning paths to multiple targets: memory involvement and planning heuristics in spatial problem solving. Psychological Research PRPF **73**(5), 644–658 (2009)
22. Wiener, J.M., Tenbrink, T.: Traveling salesman problem: The human case. KI **22**(1), 18–22 (2008)
23. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience **5**(6), 495–501 (2004)

A Problem Specifications

The following tables provide a mapping between the problem numbers used in Figures 6 and 13. The cooridnates of the problem nodes are given in the respective literature and for problems in our current game version can be downloaded at `http://www.perlentaucher.medieninformatik.uni-tuebingen.de:8888/leveldata.zip`.

Table 2 Specification of problems used in the prototype and the final version, see also Figure 6

Number	Problem name	Level in current Dataset
1	Random-3	Level 3
2	20-Nodes-12-IP[15]	Level 7
3	Random-1	Level 10
4	Random-9	Level 15
5	20-Nodes-16-IP[15]	Level 17
6	Random-0	Level 19
7	Nn-inadequate-10[21]	Level 23

Table 3 Specification of problems compared in Figure 13

Number	Problem name
MacGregor[15]	
1	10-Nodes-1-IP
2	10-Nodes-2-IP
3	10-Nodes-3-IP
4	10-Nodes-4-IP
5	10-Nodes-5-IP
6	20-Nodes-10-IP
7	20-Nodes-12-IP
8	20-Nodes-14-IP
9	20-Nodes-16-IP
10	20-Nodes-4-IP
11	20-Nodes-6-IP
12	20-Nodes-8-IP
13	Dantzig
Tak et al. [19]	
14	Circle
15	Dantzig
16	Plaisier-Tak
17	S-shaped
18	Square

B Comparison to Reference Data

Figure 13 compares the data from our last prototype with the results from the literature. For mapping of numbers in the figure and problem specifications, see Table 4.

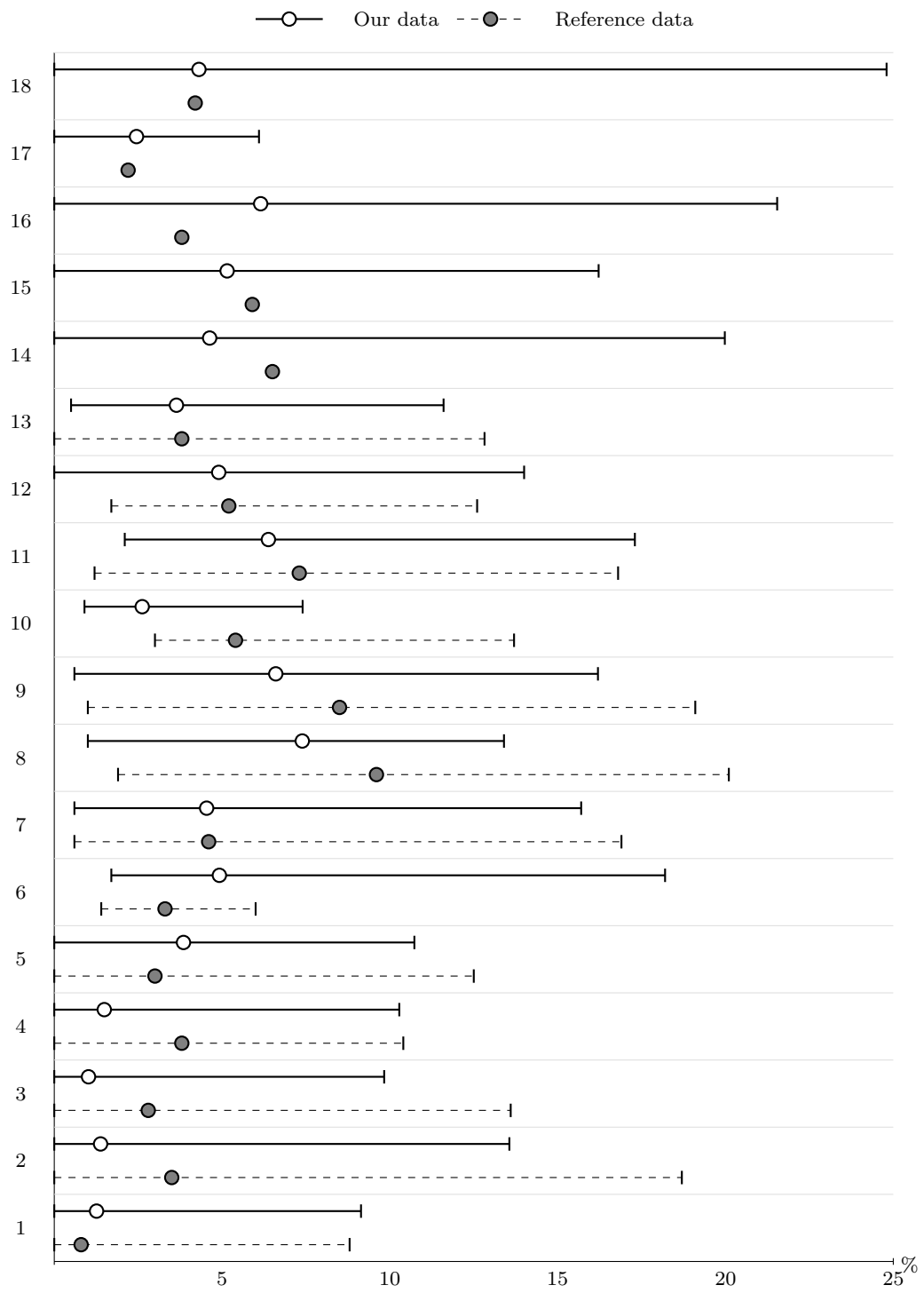


Fig. 14 The minimum, maximum and mean PAO values acquired in the prototype in comparison with the corresponding reference data