7-2012

Using Interactive Evolutionary Computation (IEC) with validated surrogate fitness functions for redistricting

Christine CHOU
National Dong Hwa University

Steven KIMBROUGH
University of Pennsylvania

John SULLIVAN-FEDOCK
University of Pennsylvania

C. Jason WOODARD
Singapore Management University, jason.woodard@olin.edu

Frederic H. MURPHY
Temple University

DOI: https://doi.org/10.1145/2330163.2330312

Published in GECCO'12: Proceedings of the 14th International Conference on Genetic and Evolutionary Computation, July 7-11, 2012, Philadelphia, Pages 1071-1078.
http://doi.org/10.1145/2330163.2330312

Using Interactive Evolutionary Computation (IEC) with Validated Surrogate Fitness Functions for Redistricting

Christine Chou
National Dong Hwa University
Department of International Business
Hualien 97401,Taiwan

Steven O. Kimbrough
University of Pennsylvania
Operations and Information Management
Philadelphia, PA 19104

John Sullivan-Fedock
University of Pennsylvania
Operations and Information Management
Philadelphia, PA 19104

C. Jason Woodard
Singapore Management University
School of Information Systems
Singapore 178902

Frederic H. Murphy
Temple University
Marketing and Supply Chain Management
Philadelphia, PA 19122

ABSTRACT

We describe a novel use of evolutionary computation to discover good districting plans for the Philadelphia City Council. We discovered 116 distinct, high quality, legally valid plans. These constitute a rich resource for stakeholders to base deliberation. This raises the issue of how to deal with large numbers of plans, especially with the aim of avoiding gerrymandering and promoting fairness. Interactive Evolutionary Computation (IEC) is a natural approach here, if practicable. The paper proposes development of Validated Surrogate Fitness (VSF) functions as a workable and generalizable form of IEC.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms

Keywords

interactive evolutionary computation, evolutionary programming, districting, compactness

1. INTRODUCTION

Typically in democracies, representatives are elected from geographic areas having roughly equal populations and this is required by law. Since population shifts are ongoing, it is necessary periodically to redraw the boundaries of electoral districts in order to maintain approximately equal populations among them. This process is called (re)districting or reapportionment. It is fraught due to the fact that there are often many legally valid ways to draw the boundaries of the districts, but some ways will favor one party or another and parties in power will normally attend to their own interests ahead of the public's.

Districts that are grossly and oddly shaped for the benefit of some party, cause, or even individual (a common occurrence) are said to be *gerrymandered.*[1] Aside from partisan griping by losers, there is widespread agreement that gerrymandering is not in the public's interest. Among other things, this has led to efforts to get the public involved by encouraging the development and publicizing of districting proposals by the general public in hopes that the presence of better plans will lead to better outcomes.

This paper describes two main contributions pertaining to this context and having wider import as well. The first has to do with a novel use of evolutionary computation in which more than 100 legally valid, high quality districting plans were generated for the Philadelphia City Council reapportionment process in response to the 2010 census. This was undertaken in conjunction with a competition in 2011 sponsored by Azavea (`http://www.azavea.com/`), a firm specializing in use of geographic data. The philosophy motivating this first contribution has been called *solution pluralism* (e.g., [11]). The notion is that decision making is often well served by providing a plurality of good solutions to support deliberation. A similar point has routinely been made in the IEC (Interactive Evolutionary Computation) literature (e.g., [3, 20]).

The second main contribution of the paper pertains to the consequences of producing a large number of high quality solutions. Electoral districts must by law be contiguous and close in population (of which more below). These properties are readily computed. In addition, the antidote to gerry-

[1]After Governor Elbridge Gerry of Massachusetts who in 1812 engineered a districting plan in which one of the districts resembled a salamander. The term "Gerry salamander" was soon shortened to "gerrymander" and the name has stuck.

mandering is generally agreed to be compact districts. The problem is that there are very many definitions of compactness, none of which is generally accepted as normative and all of which are known to have serious problems [24]. Can we at the least identify a computational measure of compactness that accords well with subjective judgments? If so, that measure could be used both to direct evolutionary (more generally, heuristic) search and to winnow the consideration set used for deliberation and subjective judgments. Our second main contribution consists of a number of experiments with subjects, the upshot of which is that the measure of compactness we employed in our evolutionary search does indeed appear to be in broad accord with the collective judgments of our experimental subjects.

In what follows, we present and discuss these two main contributions in order. The paper concludes with a summary discussion and comments about the larger import of our findings. First, however, a brief discussion of related work.

2. RELATED WORK

The *districting* problem is also known as the *zone design*, the *territory design*, and the *commercial territory design* problem, with the first term more prevalent in contexts of reapportionment of electoral districts and the latter terms more common in industrial applications such as designing school districts, police and fire districts, sales and service districts, and so on (see [2], [4], [14], [16], [17], and especially [9] for useful discussions and references).

However called, the problem arises when there are n areal (geographic) units that must be assigned to k groups (districts, zones) where a value function is optimized, subject to constraints. In electoral districting, it is common for the objective to be the compactness of the districts, subject to constraints on population size and an absolute requirement for contiguity (see discussion below). Zone design can be formulated as a variety of knapsack problem. It is known to be NP-complete [1] and very challenging in practice. In consequence, it is normally attacked heuristically, either with traditional OR methods or with metaheuristics. Oddly, very few papers have appeared using evolutionary computation for districting ([2] is an exception; see also [5]). We are not aware of any work that seeks to produce a large plurality of solutions, which is the focus of our efforts.

Interactive Evolutionary Computation may fairly be dated to the publication of Dawkins's book *The Blind Watchmaker* [6]. There he presented his Biomorph program which used an L-system grammar to generate plant-like branching arrangements. (See [13] for a recent overview of L-systems and evolutionary design.) Users could select one from about 20 images shown at once on the screen, and the program would respond with another 20 images produced by mutation on the chosen one. After a few generations of "evolution," strikingly interesting and complex individuals could be created. (Many re-implementations of the Biomorphs program can easily be found with Web searches and exercised.)

Dawkins's purpose was to demonstrate the power of natural selection. The evolutionary computation community hardly needed any convincing; very quickly Dawkins's idea was abstracted, leading to a large number of studies using human judgment to augment or replace computed function evaluation for assessing fitness. See [20] for a comprehensive review of the extensive literature as of 2001. The earlier essay by Bentley [3] also remains useful and covers somewhat different ground. Besides purely academic research, the IEC concept has spawned fascinating applications (e.g., Electric Sheep `http://electricsheep.org/` and PicBreeder `http://picbreeder.org/`; see also [19]), successful commercial design ventures (e.g., `http://www.affinnova.com/`, `http://www.natural-selection.com/`), and academic entrepreneurship (e.g., `http://mitsloan.mit.edu/vc/`).

Our idea (see below) of obtaining subjective judgments for validating a surrogate fitness function is not entirely unanticipated. Both [8] and [22] report using questionnaires to elicit information from users that directs an evolutionary algorithm in searching a design space. The number of users in each case is quite small (even equal to one) and the motivation is to reduce the burden of fatigue on the customer of the system.

A terminological point: We use the term IEC since it approximates being a standard. Other terminology includes: Aesthetic Evolutionary Artificial Life and Aesthetic Evolutionary Design (both in [3] and elsewhere); Simulated Breeding, Simulated Evolution, and Interactive Evolution (all three in [22]); Aesthetic Evolution (e.g., `http://evonet.lri.fr/eurogp2005/?page=evomusart`);[2] Collaborative Art and Evolutionary Art (e.g., `http://picbreeder.org/`).

3. DISTRICTING PROBLEMS

Ten members of the Philadelphia City Council are elected from 10 areal districts. Several additional members are elected "at large," that is without being tied to any particular district. The districting plan from the 2000 census (in force until 2015) is notoriously gerrymandered. See Figure 1. In parallel with the 10 councilmanic districts, Philadelphia has 66 wards, each of which is divided into between 10 and 50 ward divisions, of which there are more than 1300 in all (see `www.seventy.org`). These are political entities determined by the City. Each ward has one or more ward captains. These captains are associated with the political parties and so the wards are meaningful entities in terms of "neighborhood," although imperfectly so, since the wards were last defined in the mid-1990s.

There are three generally agreed criteria that districting plans should meet. Philadelphia adds a fourth.

1. Contiguity. Every point in a district must be reachable from every other point in the district. This is an absolute requirement of the law.

2. Equal populations. Districts should be approximately equal in population. For Philadelphia the courts have accepted a maximum of a 10% difference between any two City Council districts. However, any plan that is near this limit is vulnerable to court challenge. We planned for a maximum 5% difference. This is standard practice.

3. Compact. The cliché of pornography—that it can be recognized but not defined—is apt for compactness. Districts 5 and 7 in Figure 1 are badly gerrymandered and intuitively very non-compact. Pennsylvania and Philadelphia do not recognize in law any definition of compactness, nor is there any generally accepted definition [24], as noted above. The role of this concept

[2]See also `http://draves.org/evomusart05/-evomusart05draves.pdf`.

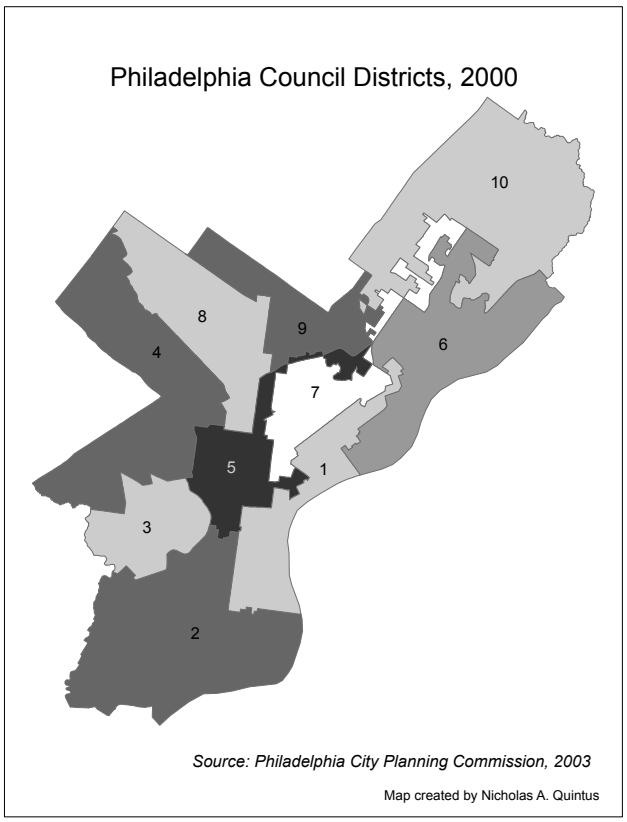

Figure 1: Current districting map for Philadelphia City Council (year 2000 census). www.seventy.org

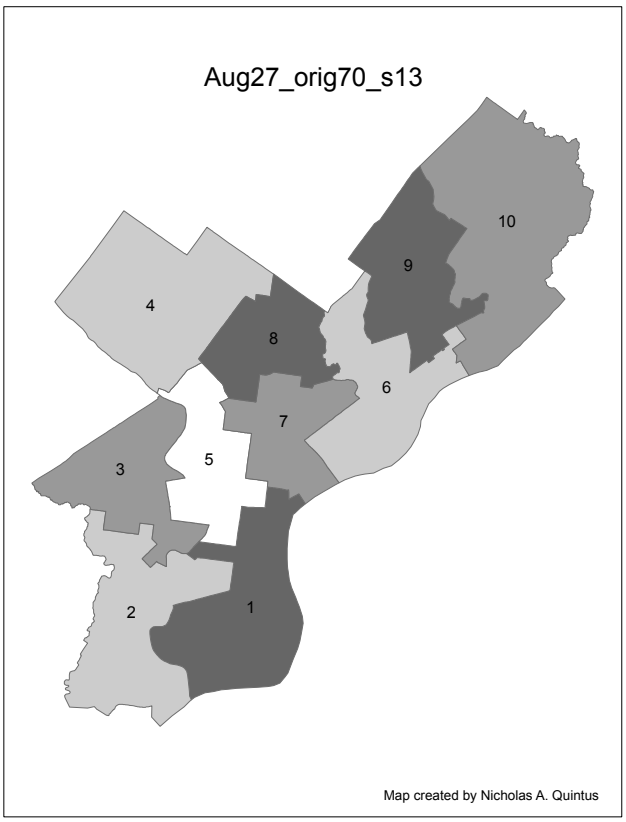

Figure 2: Example districting found by our algorithms: s13. Contiguous, populations within 5% of each other, and none of the 66 wards broken up. Based on 2010 census.

in pubic debate is nevertheless important, even if it is based on intuitive judgments.

4. Preserve neighborhoods. Philadelphia, as is often said, is a city of neighborhoods and ethnic groups. One form of gerrymandering is to divide a neighborhood into multiple districts in order to minimize the influence of the neighborhood group. District 7 in Figure 1 did this to Philadelphia's Latino population. As it happened, Philadelphia gained population in the 2010 census for the first time since the 1950 census and the group with the largest increase (about 59,000 people) was the Latinos. Hence, there was considerable pressure to de-gerrymander the 7th district. A difficulty here is that no one has other than a subjective characterization of Philadelphia's neighborhoods. The closest thing available are the wards.

Our program and in consequence our districting plans operated at the ward level. That is, each of our plans allocated each of the 66 wards to one of ten councilmanic districts. (So the raw search space was ${}_{66}C_{10} \approx 2 \times 10^{11}$.)

In preparation for programming and running our evolutionary computation, we had to acquire several forms of data. US census data, collected every 10 years, is the basis for districting and reapportionment. This data is organized by "census blocks" which are generally smaller than ward divisions and in any case defined completely independently of them. Census blocks may, and routinely do, cross ward division and ward lines. Thus, we needed to determine the population of each ward (since our program uses wards as the smallest units of area). From City of Philadelphia sources we obtained ward x–y location and ward adjacency information, allowing us to create a matrix of ward distances and a ward adjacency matrix. Finally, we obtained four districting plans with contiguous districts. Each of our runs, as will be explained, was initiated with a single one of these four original plans. The four plans in question were obtained by exact classical optimization of a heuristic for creating contiguous districts. Exact optimization of the problem proved impossible and these four plans are not of good quality, except on contiguity. We resorted to this approach after failing to obtain contiguous solutions through evolutionary computation. We note that other heuristics, such as for the maximum diversity problem, are available for creating contiguous starting points (e.g., using an iterated greedy algorithm [12, 18]).

The evolutionary algorithm we used resembles evolutionary programming (cf., `http://www.scholarpedia.org/article/Evolutionary_programming` and [7] for overviews), but has its own twists, which we now explain briefly. See the supplementary material for details: `http://opimstar.wharton.upenn.edu/~sok/phillydistricts/doc/` then `pap239s1-supplementary-material.pdf`.

Given a single contiguous districting plan (which did not need to be feasible with respect to population equality), we created the initial population by mutating the solution until

we obtained a sufficient number of new, contiguous solutions (49 new in the case of a population size, popSize, of 50). Mutation here and throughout was conducted by what we call *neighborhood mutation*. For a solution undergoing mutation, we obtain the wards in the solution that are adjacent to wards assigned to a different district. These wards may mutate to any neighboring district.

To create a new generation, each member of the current population produced 2λ offspring by mutation. (After some tuning, we set $\lambda = 3$ for the actual runs. λ offspring were produced by a higher mutation rate of 0.15 per locus; other λ by a lower mutation rate, 0.01. This produces a pool of solutions 7 times the population size.) We did not use recombination. Every member of the pool of solutions is then evaluated for fitness.

Our objective function for fitness was a measure of compactness. Recall that each ward has a pair of geographic coordinates identifying the location of its conventional center. The maximum intra-district distance is the largest distance between the stipulated centers of any two wards assigned to a given district. Our measure of compactness—which we call the *max-max-distance*—is the largest intra-district distance in a districting plan. We seek to minimize it. We handle contiguity and population size as constraints. Violations of constraints resulted in penalties on the objective. Since we initialize with entirely feasible solutions with regard to contiguity and since every individual in the current population belongs to the pool of solutions evaluated for the next generation, putting a heavy penalty on non-contiguous solutions prevents them from entering into the next generation.

Regarding population size differentials, the program variable slackAllowed (=0.025, for a 5% differential) was used to set the maximum and minimum populations permitted for any district in a districting plan. Plans that violated the maximum or minimum were penalized, although not as severely as non-contiguous plans. In consequence, all feasible solutions had a better (lower) fitness value than any infeasible solution, and all solutions infeasible on contiguity had worse values than any solution infeasible only on population sizes.

Each run was for 2000 generations and each of the four original contiguous solutions was used in 100 runs. We collected every feasible solution discovered and obtained 116 distinct such solutions in all. Figure 2 shows one such solution, one that scores in the middle on our compactness measure and that seems, to those skilled in the ways of Philadelphia, to do comparatively well in honoring the existing neighborhoods. Interestingly, district 7 is close to what was advocated by the Latino community.

4. IEC

Interactive Evolutionary Computation characteristically relies on the judgments of subjects to assess the fitnesses of the solutions encountered in runs of an evolutionary algorithm. As such, IEC has a special strength and a special weakness. Its strength is that it affords discovery of designs and solutions by evolutionary processes that otherwise could not be well directed for lack of calculable fitness functions. Search is feckless without meaningful direction.

The great weakness of IEC is, of course, the cost in labor, time, and human fatigue required to perform the fitness evaluations. This "fatigue problem" places severe constraints on population sizes and generation numbers. As Takagi notes:

> The EC population size is limited by the number of individual images that are spatially displayed on a computer monitor simultaneously or by human capacity to remember sounds or images for time-sequentially displayed individual sounds or movies. The number of EC search generations is limited by human fatigue as well, and 10 or 20 EC search generations are usually the maximum [that can be used]. [20, page 3]

Considered "the most pressing problem in IEC systems" [21], the fatigue problem of IEC could be overcome, or at least ameliorated, by using subjective judgments to test, or even to develop, a computational fitness function, which then could be used more or less in the ordinary way for evolutionary computation. We call such a function a *validated surrogate fitness (VSF) function.* It is a surrogate because it is to be used instead of human judgments. It is validated after performing satisfactorily on a sample of such judgments.

Our proposal for, and interest in, VSF functions arose out of our experience with the novel use of evolutionary computation to discover districting plans for the Philadelphia City Council, described above. In what follows we describe and discuss the experiments we did to validate a particular VSF function. The paper closes with a summary discussion and pointers towards further research, development, and application.

5. EXPERIMENTS

Our GA discovered 116 legally valid districting plans, which we rendered into color maps. We used these maps as the basis for our experiments.

For each map we had a "compactness" score, ranging 334.6 to 400.42. This compactness score for a districting plan is the maximum of the maxima of the intra-district distances, the max-max-distance. That is, there are 66 wards. A districting plan apportions them into 10 districts. According to this measure, more compact plans have lower scores; we wish to minimize the max-max-distance. The empirical question we want to address is whether subjective judgments are broadly in agreement with this measure of compactness. Ultimately, it would be very desirable to have a well-supported model that could accurately predict human judgment of compactness. That would require a project well beyond the scope of this, or indeed any single, paper. Instead, we wish to investigate whether our particular measure, used in the evolutionary computation, is a reasonable one for focusing attention on a smaller consideration set. To this end we conducted three experiments, preceded by several pre-tests, which asked subjects to rank printed maps of the districting plans discovered by our evolutionary algorithm.

In the first experiment, 39 subjects were each given 2 pairs of maps and asked for each pair to judge which map was more compact. In all 4 pairs of maps were used, randomly assigned to the various subjects. (The ordering of the maps in the pairs was randomized as well.) Each of the pairs consisted of one map randomly drawn from the top 10 maps and one from the bottom 10. The best 4 maps had compactness scores between 341 and 347, while the worst 4 had scores between 391 and 400. Table 1 presents a summary of the results.

Pair:	1	2	3	4
Low	19	17	15	13
High	0	3	4	7
p-values:	1.9073e-06	0.0013	0.0096	0.1316

Table 1: Summary of results for experiment 1

The p-values for pairs 1–3 are significant statistically, disconfirming the null hypothesis of no difference between the maps with regard to compactness. Pair 4 is not significant statistically although it is in the "right" direction, favoring the map with a lower compactness score by our measure. Its failure to be significant might be a random effect or it might indicate a more difficult choice for the subjects. On this, we note that the compactness scores for pair 4 were the closest of the four pairs examined. Further, our pre-testing (as well as our second experiment) indicated the subjects had considerable difficulty choosing between closely-scored maps. Pooling the 78 observations, the probability of getting 14 or fewer high votes given the null hypothesis of a probability of 0.5 on any single vote is 4.2911e-09.

Subjects in the first experiment were undergraduates at the University of Pennsylvania. In the second experiment we used 38 undergraduates at National Dong Hwa University in Hualien, Taiwan. For this experiment, we developed two lists of 9 maps, list A and list B. Each list had 3 maps drawn at random from the best 12 maps, 3 from the middle of the range, and 3 from the bottom. The two lists had no maps in common. Each subject was asked to rank the 9 maps in a single list; the original presentation ordering of the 9 maps was randomized for each subject. Table 2 presents a summary of the results. The vote scores in the rightmost two columns indicate the percentage of subject judgments that agree with the compactness rankings on the pair associated with the row. For example, in row 2 we find 0.7895 in the B vote column. This means that 78.95% of the subjects who ranked the B list ranked the map corresponding to 1 higher (better) than they ranked the map corresponding to 9, where the 1 and the 9 are the rankings of the maps according our compactness scores. With perfect (dis)agreement, each of the scores in the vote columns would be (0)1.

Upper	Lower	List A Vote	List B Vote
1	9	1.0000	0.8421
1	8	1.0000	0.7895
1	7	1.0000	0.8421
1	6	0.6842	0.0526
2	9	0.9474	0.9474
2	8	0.9474	0.6316
2	7	0.9474	0.8947
2	6	0.0000	0.1053
3	9	1.0000	0.8947
3	8	0.9474	0.6842
3	7	0.9474	0.8947
3	6	0.0000	0.0526
4	9	0.8947	0.9474
4	8	0.8947	0.8947
4	7	0.9474	0.8947
4	6	0.0526	0.5789

Table 2: Summary of results for experiment 2

Based on the subjects' rankings, Table 2 tabulates the votes on the top 4 maps compared to the bottom 4 (as judged by our compactness statistic) in each list. The pattern that clearly emerges is that there is very strong agreement between the subjects and our compactness scores for the top 3 versus the bottom 3 of each list. If we take a majority vote, the agreement is 100%. This holds as well for the 4th best map compared to maps 7, 8, and 9. The 1–4:6 comparison, however, is very noisy and does not agree well with our compactness scores. Overall, the table shows 32 scores of which 26 are ≥ 0.5 and by majority vote would agree with the compactness score. The probability of having 6 (=32-26) or fewer disagreements, under the null hypothesis of the probability of agreement = 0.5, is 0.00027.

We can gain additional insight into the subjective judgments by examining the correlations between the individual subject rankings and the max-max-distance scores. Table 3 presents the results. The key facts for interpreting the table are these. There were two distinct lists of 9 maps used in experiment 2, lists A and B. Nineteen subjects ranked list A and another 19 subjects ranked list B. (Subjects were randomly assigned to lists.) Subject number 0 in the table refers to the max-max-distance scores produced by the evolutionary algorithm code. Columns labeled ρ report the Spearman rank correlation coefficient between the subject's rankings and the max-max-distance scores from the program. Thus, to illustrate, for list A, the correlation between subject 3's rankings and the ranking implied by the max-max-distance scores on the A list of maps is 0.6555. For list B, the correlation between subject 3's rankings (a different individual than subject 3 for list A) and the ranking implied by the max-max-distance scores on the B list of maps is 0.5148. Finally, the columns labeled p-value report the p-values of the data in the columns immediately to their left. To illustrate, the p-value of 0.0632 for list A subject 3 indicates that the probability of obtaining a ρ-value farther from 0 than 0.6555 if the null hypothesis is true (of no linear association between the max-max-distance score rankings and the subject's rankings) is 0.0632.

By way of interpreting the results in Table 3, the ρ-values for both lists are gratifyingly high. For these subjects and these maps there is in general a very positive linear association between the subjects' rankings and the rankings from the max-max-distance scores. We note three exceptions. Subject 16 in list A and subject 12 in list B have ρ-values that are outliers in being small in their absolute values (both happen to be negative). Their announced rankings are more or less random with respect to the max-max-distance scores. This suggests that the two subjects either did not take the experiment seriously or were simply unable to discern the relevant differences among the maps. The other anomaly is subject 8 in list B, who has a highly significant linear association with the max-max-distance scores, but in the wrong direction! We suspect this subject simply was confused and presented the ranking inversely, from worst to best, instead of (as instructed) from best to worst.

On the basis of these two experiments, we have support for the hypothesis that indeed subjects broadly agree that lower max-max-distance measures compactness reasonably well. The support is far from being dispositive empirical evidence, but so far as it goes it is very encouraging. Absent additional information, in a practical setting it surely warrants focusing attention on districting plans with better (lower) max-max-distance scores.

	List A		List B	
Subject No.	ρ	p-value	ρ	p-value
0	1.0000	0.0001	1.0000	0.0000
1	0.6214	0.0826	0.8102	0.0116
2	0.5278	0.1481	0.7511	0.0246
3	0.6555	0.0632	0.5148	0.1595
4	0.5789	0.1092	0.8018	0.0127
5	0.5448	0.1339	0.8355	0.0077
6	0.5448	0.1339	0.8777	0.0033
7	0.5789	0.1092	0.4979	0.1753
8	0.5703	0.1156	-0.7680	0.0199
9	0.6895	0.0474	0.8777	0.0033
10	0.4427	0.2322	0.7764	0.0182
11	0.7406	0.0286	0.6161	0.0838
12	0.6895	0.0474	-0.0928	0.8171
13	0.5703	0.1156	0.7258	0.0323
14	0.6470	0.0673	0.5992	0.0946
15	0.5703	0.1156	0.9030	0.0018
16	-0.1617	0.6794	0.7258	0.0323
17	0.5448	0.1339	0.7342	0.0299
18	0.7491	0.0260	0.8693	0.0037
19	0.5703	0.1156	0.5570	0.1246

Table 3: Experiment 2 correlations. ρ = Spearman's rank correlation coefficient

A word or two about the conduct of the first two experiments. First, the subjects were told they would receive a monetary reward based on their performance. Second, their performance would be scored along the lines of a "beauty contest." For example, in experiment 1 the following instructions were read to the students:

> Your job in this experiment is to view 2 pairs of different, legally valid districting plans for the City of Philadelphia, Pennsylvania, USA. You are to rank each of the two pairs in regard to apparent fairness or compactness. Your performance will be scored by comparing your judgments with the judgments of other subjects in the experiment; the more you agree with the judgments of your peers, the higher your score will be. You will be rewarded according to the score you achieve.
>
> To illustrate, the figure below is a representative example of a legally valid districting plan for Philadelphia. Each of the 10 required districts is indicated by a single color. In the actual experiment, you will be given 2 pairs of such plans, one for question 1 and one for question 2, and asked to decide which of the two plans is, in your view of others' views, better than the other with regard to compactness/fairness.
>
> ...I will tally the class votes. Each student will get $1 for each choice that agrees with the majority. So, you can win at most $2 and at least nothing.

Third, because pre-tests indicated we needed to make the idea of a "beauty contest" game salient for the subjects, we first introduced the concept by having them play another beauty contest game. Here are the instructions we used for both experiments (although the second experiment was translated into Mandarin).

> In part 2 of this exercise we will be conducting an experiment in which you will be asked to make a series of judgments. Depending on how well you do, as we will explain, you could obtain a monetary reward for your performance. This applies to part 2. In part 1, where we are now, we will conduct an experiment that resembles the part 2 experiment in certain ways. You will be asked to make some choices as if you were to be rewarded based on your performance, but we will not actually give out any rewards in part 1. The rewards will come only in part 2. The purpose of part 1 is to help you understand your task in part 2.
>
> Now to the part 1 exercise, which is very brief.
>
> Here is the situation. You are asked to pick one number between 1 and 10, that is, 1 or 2 or 3 or ...or 10. So you have 10 choices possible. You are to choose your number without communicating with any other person in the room. After everyone has chosen, the results will be tallied. NOW THE IMAGINARY PART. Your score for your choice will simply be the number of people in the room who made the same choice. So, for example, if 3 people pick 8, then each of them gets 3 units of reward. So, obviously, when you make your choice you want to figure out as best you can what the other people in the room will choose. Remember: no communication.
>
> Are there any questions? [If so, answer them.]
>
> OK, now since this is for practice only, I am asking you to think for a minute and make your choice. ...Ready? OK, let's see by a show of hands what the choices were. [Go to the blackboard and record the number of 1s, 2s, etc. Then discuss briefly why people chose as they did.]
>
> Good. Now this kind of game (and it is a game in the sense of game theory) has been called a "beauty contest" after real contests in which participants were asked to select who they thought the other participants would see as the most beautiful girl.
>
> In part 2, we will also have a kind of beauty contest, like the one we just did, although it will be for a less frivolous, more serious purpose. Questions?
>
> OK, then, on to part 2.

We ran a third experiment using the crowdsourcing tool Amazon Mechanical Turk (AMT). Our aim was to replicate experiment 1 (approximately). If AMT can reliably be used for IEC purposes, and especially for developing VSF functions, the approach we are proposing and exploring here would seem to offer the prospect of very wide application and utility.

We felt that our instructions would have to be very brief on AMT. We limited them to the following message, which (we hope) succinctly communicates the beauty contest aspect of the task.

> Your job in this experiment is to view 4 pairs of different, legally valid districting plans for the

> City of Philadelphia, Pennsylvania, USA. You are to rank each of the pairs in regard to apparent compactness. Your judgments will be compared with the judgments of other subjects in the experiment. You will receive a $0.10 bonus if your selection matches the majority, so try to pick the maps that you think other people are most likely to pick.

The same random pairs of districts were used as in experiment 1. Each of the 4 pairs was included in every survey, so participants were each asked to make judgments on all 4 cases (unlike experiment 1, where each subject judged two pairs of maps). Subjects were required to have an approval rate above 95% on AMT. Participants were paid $0.05 per task (4 in all) for participating and rewarded with a $0.10 bonus for selecting the most popular districting schemes (and thus could earn up to $0.40 in bonus). The rates were based on the predicted time needed to complete the survey. Subjects were prohibited from taking the test more than once. In all, 137 surveys were completed. The results are summarized in Table 4, which is directly analogous to Table 1. These results broadly agree with experiment 1. In both cases, a majority in all four tasks agrees with the max-max-distance measure regarding which of the two compared maps has more compact districts. In both cases, 3 of the 4 individual tasks have statistically significant outcomes on their own. And in both cases the aggregate voting is highly significant statistically. For experiment 3, there were 321 "successes" in 548 trials. Under the null hypothesis and assuming a binomial distribution the probability of getting this many or fewer "failures" is 3.4199e-05 (= binocdf(548-321,548,0.5) using MATLAB notation).

Pair:	1	2	3	4
Low	83	73	79	86
High	54	64	58	51
p-values:	0.0082	0.2472	0.0436	0.0018

Table 4: Summary of results for experiment 3. (Compare with Table 1.)

Also, in all 8 of the tasks in experiments 1 and 3 the subjects voted in accord with the max-max-distance criterion. Under the null hypothesis (of a very low power model) the probability of this happening is $\frac{1}{2}^8 = 0.0039$.

The average time taken by a subject to complete the 4 tasks was 143 seconds. Subjects who got all 4 tasks "correct" spent an average of 192 seconds. The 3s spent 133 seconds on average, the 2s 129 seconds, the 1s 124 seconds, and the 0s (those who were "wrong" on all 4) 125 seconds. Very interestingly, overall, more time spent on the tasks does not associate with better than average performance in any straightforward way.

It is by now a well-established practice to run experiments using subjects doing HITs (Human Intelligence Tasks, AMT's jargon) on Amazon Mechanical Turk [15]. Even so, there are continuing worries about the representativeness of the subject population and the associated noise and biases. It is therefore especially gratifying to find the AMT results to be broadly in accord with a more conventional human subject experiment. The tantalizing prospect is that large scale crowdsourcing (as on AMT) can be used to discover and validate surrogate fitness functions, thereby greatly expanding the scope of application of interactive evolutionary computation.

6. DISCUSSION

There are two main contributions made by this paper. First, we have used evolutionary computation to generate a significant number of high quality solutions (solutions that are legally valid and in fact more compact than the ones the political process settled on, however measured) for a districting problem, in which it is difficult to discover even a few very good solutions. Having such a plurality of good solutions is, we believe, potentially of great value for supporting deliberation, both in the particular case of reapportionment and in general. Regarding apportionment, we would extend the vision of Azavea (and many others) that foresees an empowered public weighing in on redistricting decisions by presenting and arguing for solutions more in the public interest. Why not use the ability to generate a large number of good solutions to *constrain* the decision process to a suitably characterized pool, each member of which is reasonably compact and meets the legal requirements on contiguity and population size? Set a loose requirement on compactness, perhaps as measured by several different rules. Use evolutionary computation, or indeed any other method, to generate a large consideration set of objectively good quality solutions. Let anyone contribute to this pool, but limit the consideration set to the pool. Confine the discussion to which member of the pool is to be chosen. This leaves room for a modicum of politics and for incorporation of other features not incorporated into the computational models, and it secures choice among objectively good solutions. A regime on this order (further details are needed, of course) could in fact be implemented by agreement of non-partisan districting boards, where they exist. Other applications—such as design of sales districts—might follow a similar procedure informally. In his review of compactness measures for districting, Young concludes that "compactness is such a hazy and ill-defined concept that it seems impossible to apply it, in any rigorous sense, to matters of law" [24, page 113]. The approach we have sketched may well afford keeping the baby without the bath water. If one or a pool of compactness measures can be empirically validated, then surely the move we have indicated of focusing discussion on the better-scoring solutions could be used at the least to shift the burden of argument onto those who would advocate significant departures. And very many extant districting plans are in fact in this category. (Compare the two maps above.) This said, it must be admitted that this proposal is here only speculative. Its proper investigation is apt for future research.

Generalizing further, we observe that population-based metaheuristics, among them evolutionary computation, are well suited to the task of finding many good solutions. For starters, one needs merely to capture them in the normal run of events. Such pools of good solutions will, we believe, often be useful for sensitivity and robustness analysis as well as in situations such as zone design where aspects of the model (compactness in the present case) are imprecisely known. This is a subject for future research (but see [10] for a treatment of robustness).

Our second main contribution has to do with using subjective judgment to validate a fitness function in a difficult context. IEC is used in just such circumstances for lack of

good automated alternatives. Further, if per our first contribution decision makers and stakeholders are dealing with large numbers of good solutions for which fitness functions are problematic, what can be done to assist them in selecting the better solutions for focused attention? Our proposal is to use a population of subjects to evaluate members of the consideration set of solutions and then to use their judgments collectively to validate (and if need be develop) a computational fitness function. To the best of our knowledge we are the first to do this with regard to compactness in the context of districting. There is no reason we can see to limit the approach to this context. We note that a formula has been developed for artwork (paintings) based on surveys [23]. Although this effort was in large part facetious, we think it does convey a salutatory message. If not art, then perhaps new product development for consumer goods might learn from these experiences.

7. ACKNOWLEDGMENTS

Thanks to Azavea for sponsoring the contest and providing us with key information. Thanks to Ram Gopalan and Nicholas Quintus for optimization work and GIS work. Thanks to Knowledge@Wharton for a concise, accurate story, `http://knowledge.wharton.upenn.edu/article/2850.cfm`. Our project repository is at `http://opimstar.wharton.upenn.edu/~sok/phillydistricts/`.

8. REFERENCES

[1] M. Altman. Is automation the answer: The computational complexity of automated redistricting. *Rutgers Computer and Technology Law Journal*, 23(1):81–142, 1997.

[2] F. Bação, V. Lobo, and M. Painho. Applying genetic algorithms to zone design. *Soft Computing*, 9:341–348, 2005. DOI 10.1007/s00500-004-0413-4.

[3] P. Bentley. Aspects of evolutionary design by computers. `http://arxiv.org/html/cs/9809049/dss5.html`, 23 September 1998. Intelligent Systems Group, Department of Computer Science, University College London, Gower St., London WC1E 6BT, UK. Accessed 20120114.

[4] B. Bozkaya, E. Erkut, and G. Laporte. A tabu search heuristic and adaptive memory procedure for political districting. *European Journal of Operational Research*, 144(1):12 – 26, 2003.

[5] D. Datta, J. R. Figueira, C. M. Fonseca, and F. Tavares-Pereira. Graph partitioning through a multi-objective evolutionary algorithm: A preliminary study. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, GECCO '08, pages 625–632, New York, NY, USA, 2008. ACM.

[6] R. Dawkins. *The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe without Design*. W. W. Norton & Company, New York, NY, 1985.

[7] A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Springer, Berlin, Germany, 2003.

[8] F.-C. Hsu and P. Huang. Providing an appropriate search space to solve the fatigue problem in interactive evolutionary computation. *New Generation Computing*, 23(2):115 – 127, 2005.

[9] J. Kalcsics, S. Nickel, and M. Schröder. Towards a unified territorial design approach—Applications, algorithms and GIS integration. *Sociedad de Estadística e Investigación Operativa Top*, 13(1):1–74, 2005.

[10] S. O. Kimbrough, A. Kuo, and H. C. LAU. Finding robust-under-risk solutions for flowshop scheduling. In *MIC 2011: The IX Metaheuristics International Conference*, Udine, Italy, 25–28 July 2011.

[11] S. O. Kimbrough, A. Kuo, H. C. LAU, F. H. Murphy, and D. H. Wood. Solution pluralism and metaheuristics. In *MIC 2011: The IX Metaheuristics International Conference*, Udine, Italy, 25–28 July 2011.

[12] M. Lozano, D. Molina, and C. García-Martínez. Iterated greedy for the maximum diversity problem. *European Journal of Operational Research*, 214:31–38, 2011.

[13] J. McCormack. Evolutionary L-systems. In P. F. Hingston, L. C. Barone, and Z. Michalewicz, editors, *Design by Evolution: Advances in Evolutionary Theory*, Natural Computing Series, pages 169–196. Springer, Berlin, Germany, 2010.

[14] F. H. Murphy, S. W. Hess, and C. G. Wong-Martinez. Politics. In S. Gass, editor, *Encyclopedia of OR*. 2012.

[15] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–19, August 2010.

[16] R. Z. Ríos-Mercado and E. Fernández. A reactive GRASP for a commercial territory design problem with multiple balancing requirements. *Computers & Operations Research*, 36:755–776, 2009.

[17] F. Ricca and B. Simeone. Local search algorithms for political districting. *European Journal of Operational Research*, 189:1409–1426, 2008.

[18] R. Ruiz and T. Stützle. A simple and effective iterated greedy algorithm for the permutation flowshop scheduling problem. *European Journal of Operational Research*, 177:2033–2049, 2007.

[19] K. O. Stanley and R. Miikkulainen. A taxonomy for artificial embryogeny. *Artificial Life*, 9:93–130, 2003.

[20] H. Takagi. Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the IEEE*, 89(9):1275–1296, September 2001.

[21] H. Takagi and H. Iba. Interactive evolutionary computation. *New Generation Computing*, 23(2):113 – 114, 2005.

[22] T. Terano and Y. Ishino. Knowledge acquisition from questionnaire data using simulated breeding and inductive learning methods. *Expert Systems with Applications*, 11(4):507 – 518, 1996. The Third World Congress on Expert Systems.

[23] J. Wypijewski, editor. *Painting by Numbers: Komar and Melamid's Scientific Guide to Art*. Farrar Straus Giroux, New York, NY, 1997.

[24] H. P. Young. Measuring the compactness of legislative districts. *Legislative Studies Quarterly*, 13(1):105–115, February 1988.