Region Estimation for Dialect Features Using a Cellular Automaton

Ilkka Juuso
Department of Computer Science and Engineering, Erkki Koiso-Kanttilan katu 3, FIN-90014 University of Oulu, Finland
ilkka.juuso@ee.oulu.fi

William A. Kretzschmar, Jr.
Department of English, Park 317, University of Georgia, Athens, GA 30602, USA
kretzsch@uga.edu

Abstract

An issue in dialect research has been how to make generalizations from survey data about where some dialect feature might be found. Pre-computational methods included drawing isoglosses or using shadings to indicate areas where an analyst expected a feature to be found. The use of computers allowed for faster plotting of locations where any given feature had been elicited, and also allowed for the use of statistical techniques from technical geography to estimate regions where particular features might be found. However, using the computer did not make the analysis less subjective than isoglosses, and statistical methods from technical geography have turned out to be limited in use. We have prepared a cellular automaton (CA) for use with data collected for the Linguistic Atlas Project that can address the problems involved in this type of data visualization. The CA plots the locations where survey data was elicited, and then through the application of rules creates an estimate of the spatial distributions of selected features. The application of simple rules allows the CA to create objective and reproducible estimates based on the data it was given, without the use of statistical methods.

1 Introduction

A continuous issue in dialect research has been how to make a reasonable generalization from survey data about where some dialect feature might be found in the survey area. Pre-computational methods included drawing isoglosses or using shadings to indicate areas where an analyst expected a feature to be found. The use of computers allowed for faster plotting of locations where any given feature had been elicited, and also allowed for the use of statistical techniques from technical geography to estimate regions where particular features might be found. However, using the computer as a cartographic tool did not make the analysis necessarily less subjective than old-fashioned isoglosses, and statistical methods from technical geography have turned out to be limited in use, sometimes through problems in matching statistics to the data, in other cases from the need for advanced statistical training.

We have prepared a cellular automaton (CA) for use with data collected for the Linguistic Atlas of the Middle and South Atlantic States (www.lap.uga.edu) which can address the problems involved in this type of data visualization. The CA plots the actual locations where survey data was elicited, and then creates an estimate of the spatial distributions of the variants for any selected features. The spatial estimates created by the CA are the result of the application of simple rules under which an elicited feature may be retained at a location and under which it may be adopted by a neighboring location. The task of identifying the regions for each feature is thereby handed over to the CA, which creates objective and reproducible estimates based on the data and the rules it was given. Moreover, the estimates are created without the use of statistical estimation methods, and so they require neither advanced statistical training nor a match of the data for a particular method (such as normally distributed data for many statistics).

2 Visualizing regions in language

Dialectology began in the late nineteenth century, when the Neogrammarians expected that dialect surveys would support their theories about systematic language change. However, as Walter Mitzka said of Georg Wenker, the first modern dialectologist (1943, cited and trans. in Kretzschmar 2002: 81):

> [Wenker] reported in 1885 before the Giessen conference of philologists that, "I lived in the fair and calming conviction that these [linguistic] features must completely or nearly completely go together. Each assumption turned out soon enough to be utterly mistaken: the boundaries of the contemplated features stubbornly took their own way and often crossed each other."

These boundaries were isoglosses, lines drawn on a map to indicate the limit of occurrence of a dialect feature. Kurath's Word Geography (1949) is the best example of this traditional approach; Kurath offers both single-feature maps and maps of aggregated data that are intended to show dialect regions through the relative conjunction of isoglosses. As Schneider (1988) and Kretzschmar (1992a, 2003) explain, this

qualitative method is subjective and cannot be validated except by recourse to the researcher's judgment. Nonetheless, isoglosses are still used in professional publications (e.g. in Labov et al 2006), and many professional linguists and members of the public still expect that language variation will sort itself out into particular regions, both for separate features and for aggregated data.

New methods of visualization have come with the advance of computer technology (for a review, see Kretzschmar 2014). Simple plots of dialect features have been available online interactively since at least 1996 (now, e.g., at www.lap.uga.edu, sv Old Site, LAMSAS). As reported in Kretzschmar 1992b and 1996, computer plotting increased the speed for making maps 250-fold from the six hours required formerly for hand plotting, which permitted analysts to make a great many more maps and so observe a great many more distributions of dialect features. As computers improved, so did software for Geographic Information Systems (GIS) which could be used for technical geography as well as for cartography. The use of statistics permitted new kinds of visualizations, such as the use of spatial autocorrelation (Lee and Kretzschmar 1993) and density estimation (Kretzschmar and Light 1996, Kretzschmar 1996). Statistical methods allowed for maps of single features that were objective and reproducible, but these maps did not readily support the idea of discrete regions in which the features occurred. Instead, individual features were seen to occur in clusters of locations across a region, and features showed variable likelihood of elicitation within the clusters. These statistical methods did not catch on among language variationists, perhaps because they required advanced statistics (but see Grieve et al. 2013) or because reading them seemed more complicated than previous isoglosses.

Besides generalizations for single features, computers permitted the creation of generalizations from large amounts of aggregated data. John Nerbonne pioneered the creation of maps from aggregations of Dutch dialect materials (Nerbonne et al. 1996), and with Peter Kleiweg later applied his method to more extensive American dialect materials (see www.let.rug.nl/kleiweg/L04/; Nerbonne and Kleiweg 2003). The current implementation of this line of research is called Gabmap (Nerbonne et al. 2011). Nerbonne and Kleiweg have used statistical methods in conjunction with string edit distances and multidimensional scaling methods (MDS) to create shaded areas of similar usage, which they then compare to traditional

notions of dialect regions. Kretzschmar and Thill developed the use of neural networks and self-organizing maps (SOM) to investigate aggregated data (Thill et al. 2008), which created numerous sets of weighted collections of features which could be associated with particular American survey speakers and thus make a collection of dialect maps. While these maps were interpretable in terms of traditional dialect regions, Kretzschmar has shown that the neural-network analysis actually produced a ranked list of speaker collections, not a set of dialect regions (Kretzschmar 2008). Analysis of aggregated data does not resolve the problem of generalization of dialect regions, because its statistical results remain interpretable only by the analyst's judgment in comparison with earlier assertions of dialect areas. Both MDS and SOM create what appear to be many regions, none of which exactly match previous analyses but which nonetheless evoke an imaginative response from people who look at the maps. This is still the case for the most recent statistical process which generates dialect maps, by Joshua Katz (2013; see www4.ncsu.edu/~jakatz2/project-dialect.html). Katz applied statistics to user-contributed dialect feature data collected by Bert Vaux in order to create heat maps, apparently without reference to earlier research on mapping aggregated dialect data. While Katz's maps have been popular online, they remain highly subjective in interpretation -- which perhaps accounts for their popularity.

3 Cellular automata

In this paper we present a cellular automaton (CA) that can address the problems of advanced statistics and subjective analysis, while still preserving the popular idea that dialect features should belong to distinct or even discrete areas. A CA is a particular kind of computer model (see Holland 1998 for a general outline and Sarkar 2000 for a more mathematical survey of CAs than provided here) consisting of a grid of autonomous but interacting cells that, as a system involving a large number of cells, is particularly well-suited to the study of phenomena with a strong spatial component. For our purposes here each cell relates to a single speaker on a map. The concept of a CA was first introduced in the 1940s by Stanislaw Ulam and John von Neumann (see von Neumann and Burks 1966), but only received wider interest after John Conway’s Game of Life simulation became available in the 1970s (Gardner 1970).

Since then CAs have been applied to the study of a multitude of physical, chemical, biological, and mathematical phenomena from river systems (Topa 2006) and traffic flows (Evans, Rajewsky, and Speer 1999) to the generation of random numbers (Tomassini, Sipper, and Perrenoud 2000), to name but a few examples.

In the field of language variation studies we are only aware of three earlier descriptions of CAs: Kretzschmar 1992a described the operation of a CA for the purpose of language diffusion studies, Keller 1994 reported unpublished work on an early CA that he compared to linguistic distributions, and Grieve, Speelman, and Geeraerts 2010 implemented a CA but without specific discussion of how its operation illustrated language interaction. Kretzschmar 1992a described the rule-based process from CAs as an alternative to the use of isoglosses but his experiments were not automated. Keller (1994: 100-101) reports the work of Jules Levin from an otherwise unpublished 1988 mimeograph from the University of California-Riverside. Levin had created a small (55 x 55 cells) CA, in which he randomly seeded black and white cells. Keller reports Levin's rules as having the target cell stay the same if all of its neighbors had the same status, and that a white cell had a 51% chance of staying white if four of its neighbors were white and four were black; this process produced clustered results when run over many iterations. Keller quotes Levin (1994: 101) as saying that "I regard this as only a very primitive and abstract preliminary model that hopefully mimics linguistic interaction." Thus our use of a CA is the first that we know about that uses real survey data and performs a specific task, that of creating an areal generalization from relatively sparse survey data.

In general, CAs can be constructed in one dimension, a line of locations (most famously discussed in Wolfram 2002), or in two dimensions as in our simulation, usually in a square matrix (more dimensions and other matrix types are possible but unusual; see Torrens and Benensen 2005 for examples). Given a starting state of the matrix, simple rules are applied at each iteration, the status of all the cells in the matrix is updated according to the rules, and further iterations can then continue as long as need be. As shown in Figure 1, the green cell in the middle (represented in grayscale in print) represents the location currently making a decision whether it should adopt or maintain some linguistic variant, and

the black cells represent its five neighbors that already have the variant. The uncolored cells do not have the variant. A rule would specify the number of neighboring cells that need to have the variant for the middle cell to adopt it if it does not yet have the variant, and the number of neighboring cells that need to have the variant for the middle cell to keep the variant if it already has it. The rules themselves vary depending on the purpose of the CA. In the Game of Life rules, for example, a cell with two or three live neighbors out of a neighborhood of eight surrounding cells (called the Moore neighborhood) will stay alive, and a dead cell with exactly three of its eight neighbors alive will become alive itself. Many other rule sets are possible; Wolfram 2002 documents them all for one-dimensional CAs. Most rule sets turn all cells in a matrix on after just one or a few iterations, or turn all cells off very quickly, while other rules create alternating on/off patterns. Some rule sets create chaotic patterns that, if they do repeat themselves, do so after very long cycles of iterations. A few rule sets create complex behavior, where stable patterns appear (see Kretzschmar, Juuso, and Bailey IP).

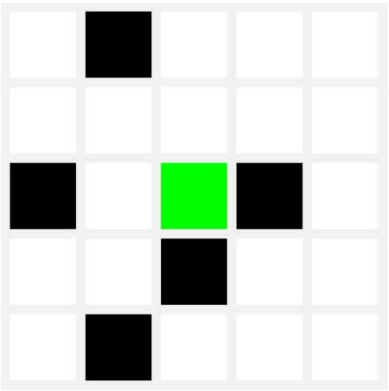

Figure 1 A target cell in its neighborhood within a CA

In this paper we investigate the possibility of using a CA to make reasonable estimates on the distribution of dialect feature variants, based on real survey data. To do this we employ a versatile CA we have implemented in Javascript for use inside HTML5 browsers, such as Google Chrome. The CA is seeded with survey data from the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS),

which was also used by Nerbonne and Kleiweg and for several of the visualization projects already noted. The CA is designed as a square matrix of user-definable proportions that forms a grid of geocoded cells to cover the whole survey area. In the work reported here we have set the CA to a size of 151 x 151, which gives us a grid of 22801 cells. The geocoded LAMSAS survey data is projected onto this grid yielding 1162 cells with real data on the linguistic features elicited at these locations as shown in Figure 2, where the black cells represent locations where one or more variants were elicited in the survey, white cells represent locations in-between survey locations, and grey cells represent locations outside the survey region. Some padding is introduced around the edges of the survey area, a layer of three cells, which addresses the problem known in GIS as "edge effects." In all, we have 8610 active cells inside the survey region, out of which 1162 cells, or 13.5%, have real survey data. Figures 3 and 4 show two examples of elicited linguistic features: Figure 3 shows the locations where LAMSAS speakers said *thundershower*, while Figure 4 shows the locations where LAMSAS speakers said *thundercloud*, each a different response to question L6#2 of the survey about what speakers called a thunderstorm.

Figure 2 Original speaker locations in LAMSAS

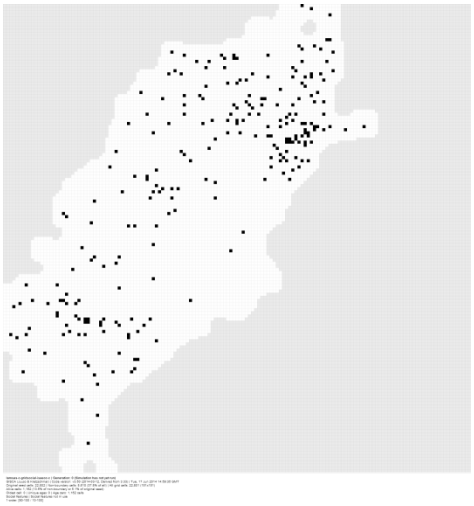

Figure 3 Locations where *thundershower* was elicited in LAMSAS

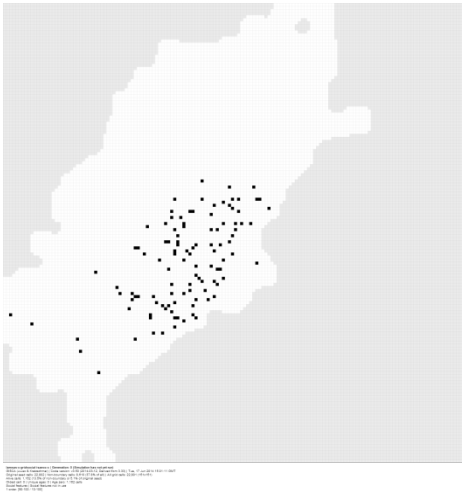

Figure 4 Locations where *thundercloud* was elicited in LAMSAS

The two responses shown in Figures 3 and 4 represent just a small sample of the 109 different variants that were recorded for this question during the survey and that we have loaded into the CA at any one time. Just a few of the variants are very common while most are rare, which creates a nonlinear profile for the ranked list of the variant counts. Out of the 109 variants, 63 variants only occur once in the data, while the maximum number of occurrences for any variant is 759 (for *thunderstorm*). In total, the 1162 informants gave 1901 responses to the question, which results in an average of 1.6 variants per informant. Figure 5 shows the distribution of speakers who gave multiple responses for this question, with darker

locations having more responses. It should be noted that speakers surveyed in Georgia, South Carolina, and upstate New York by Raven McDavid tended to give more responses per question than most speakers surveyed elsewhere, usually by Guy Lowman, so that these areas have on average more variants per cell than elsewhere. This fieldworker difference has caused Nerbonne and his associates to manage the data set, e.g. by including only speakers surveyed by Lowman in their statistical processing.

Figure 5 Multiple responses per speaker (darker locations have more responses)

3.1 Creating a region estimate

Given the locations where a variant was elicited in the LAMSAS survey, we wanted to create an estimate of the area in which the variant might be elicited among all 8610 active cells in the matrix, not just at the locations from the survey. This type of estimation is akin to how isoglosses have traditionally been drawn, but instead of attempting to draw exclusive borders between occurrences of variants we ask the CA to plot regions where neighboring speakers have a high level of agreement, represented by the rules regarding which variant or variants any cell will adopt or maintain. In this way, the CA allows us to turn the classic question of where a line between variants should be drawn on a map into one of defining what is considered to be an appropriate level of agreement. The rules that each cell uses to adopt or to maintain

a variant operate mechanically, and these mechanical interactions between adjacent locations create a pattern after the CA has run for several iterations that may look like regions defined by isoglosses but arise completely objectively. We are interested here in output from the CA that quickly resolves into a set of regions, not in chaotic or complex behavior as documented in Wolfram 2002 and Kretzschmar, Juuso and Bailey IP.[1]

The definition of the rules, the neighborhood influence scenario, is critical to the outcome produced by the CA as it, along with the original data loaded into the CA, determines the patterns that will form during the running of the CA. A scenario is defined by three factors: the conditions for variant adoption and for variant maintenance, and by the size of the neighborhood tallied for each decision. The conditions can be expressed in terms of the number of neighbors or, as we do here, the proportion of neighbors supporting a specific variant. In our scenarios, the cell neighborhood is defined as the extended neighborhood consisting of both the first and second order neighbors around each location (N=2). The choice of using the extended neighborhood of up to 24 neighbors (all cells within a distance of two cells, or N=2) instead of just the eight immediate neighbors of the Moore neighborhood (all cells within a distance of one cell, or N=1) was made to allow the CA to overcome the relatively sparse distribution of the locations in survey data. It is important to note that, because not all of the cells have data on variant use in the beginning, the number of neighbors available for each cell will vary across the CA and may be anything between 0 and 24. The percentage-based rules are always related to the number of available neighbors, and therefore also allow smaller clusters to grow if unopposed. Unlike in the Game of Life simulation, for example, where the adoption and maintenance rules have both a lower and an upper limit, here we fix the upper limit at 100% of the neighbors. In other words, all the rules reported on here set the minimum level of support for a variant to be either adopted or maintained, and assume that any higher level of support is also equally acceptable.

[1] The CA is capable of using other rule sets and neighborhood sizes, some of which produce the other output types described by Wolfram 2002, including chaotic and complex behavior.

Depending on how we set the required level of agreement, e.g. a simple majority or a two-thirds majority, the CA will produce differing estimates of where the regions may be contiguous, have voids in between, overlap, or contain individual locations that support multiple variants. Figures 7, 8, and 9 show three different estimates, each varying the rules to increase the local level of agreement applied over the initial data locations. Figure 6 shows the starting position for the thunderstorm item, where actual responses are shown in separate colors (in grayscale in the print version), and locations with multiple responses are shown in black.

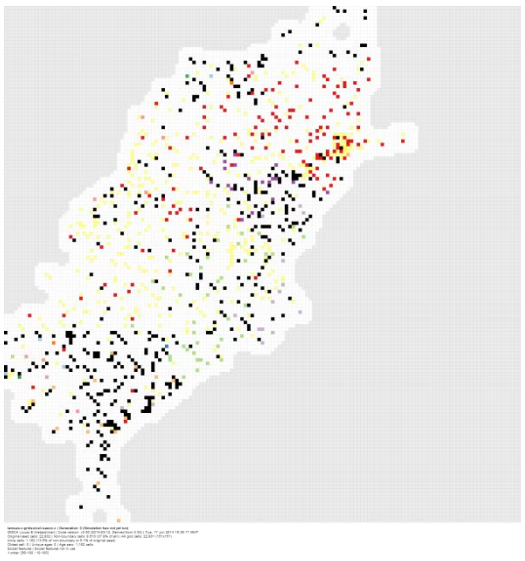

Figure 6 Starting position for thunderstorm item, with different responses in different colors and multiple responses in black

Figure 7 represents a conservative rule requiring at least 90% of the neighbors to support a variant in order for it to be adopted by a cell and only 10% or more of the neighbors to support a variant for it to be maintained.

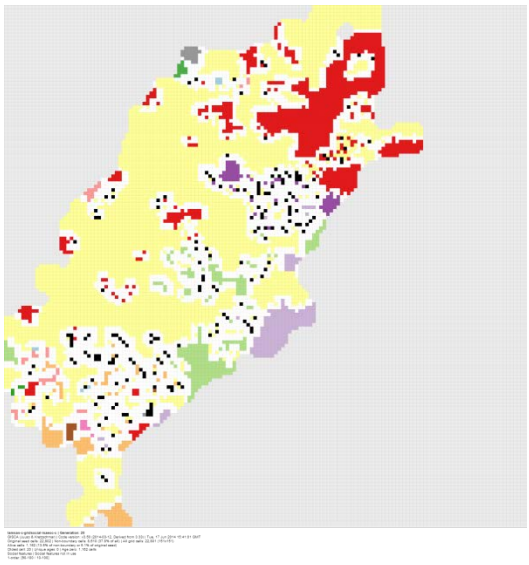

Figure 7 R9010-N2 (20 iterations)

In discussions of dialect feature boundaries, areas with black dots would be called transitional since they preserve dialect variants from two or more adjoining generalized regions. In Figure 7, eighteen variants out of the original 109 create distinct clusters after locations with missing data were removed (regions are represented in grayscale in the print version); the eight largest regions are *thunderstorm* (yellow), *thundershower* (red), *thunder gust* (purple), *thundersquall* (lilac), *thundercloud* (light green), *storm* (peach), *electric storm* (pink), and *thunder and lightning storm* (dark green). Figure 8 shows a plot for a rule that requires at least 75% support for adoption and 25% support for maintenance, resulting in a smoother pattern to emerge with very few black locations with more than one variant. The colors of the largest regions remain the same as in Figure 7.

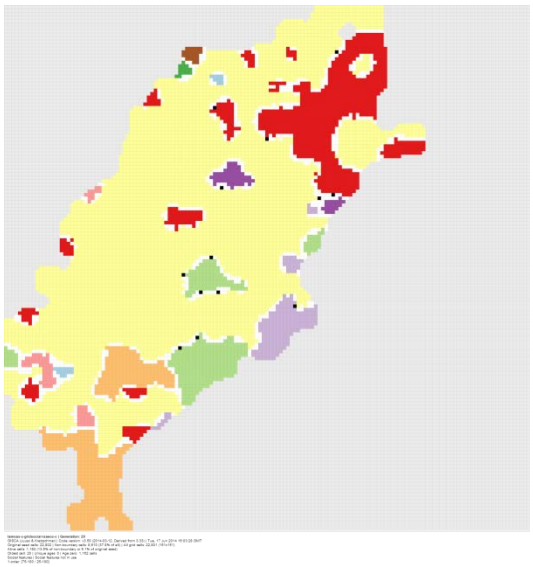

Figure 8 R7525-N2 (20 iterations)

Figure 9 modifies the rules to require at least 60% support for adoption and 40% support for maintenance, creating another estimate across the survey region with fewer small clusters and fewer black locations with multiple variants. It is important to note, however, that Figure 9 is not a derivative of Figure 8, but instead the answer to a different scenario, as is evident from comparing the island formations of the southern regions. We have found that a maintenance value of 25% or more limits the number of locations with multiple variants, creating an array of self-contained regions.

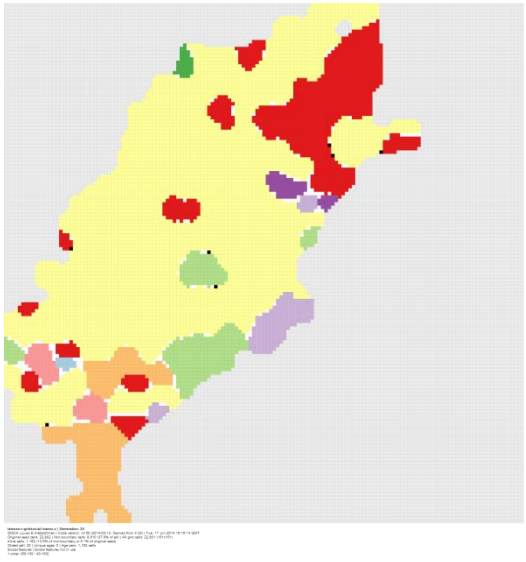

Figure 9 R6040-N2 (20 iterations)

Figures 10, 11, and 12 show what the CA has done after ten iterations for the dialect features of the thunderstorm question, using the R6040-N2 rule set. These visualizations show the original locations where a dialect feature was elicited in either black or red (black or gray in the print version), black for locations that are still alive after ten iterations, red for locations that are no longer alive after ten iterations. The green areas (gray shaded areas in the print version) are locations that form the generalization made by applying the CA rule: green locations have been turned on over the course of ten iterations. No further changes will occur for most of the thunderstorm variants after ten iterations except for the word *thunderstorm* itself, which continues to develop slightly until it becomes fixed by twenty-five iterations. In some cases the CA reaches a mostly fixed state, after which only a few cells located at borders between adjacent variant regions blink, crossing back and forth between the variant regions. This small variability in the number of iterations at which the regions become fixed has led us to adopt twenty iterations as the standard number at which we observe the maps. As the figures indicate, the live black original locations are all located inside of the green regions, while the red original locations are all outside of the green generalizations.

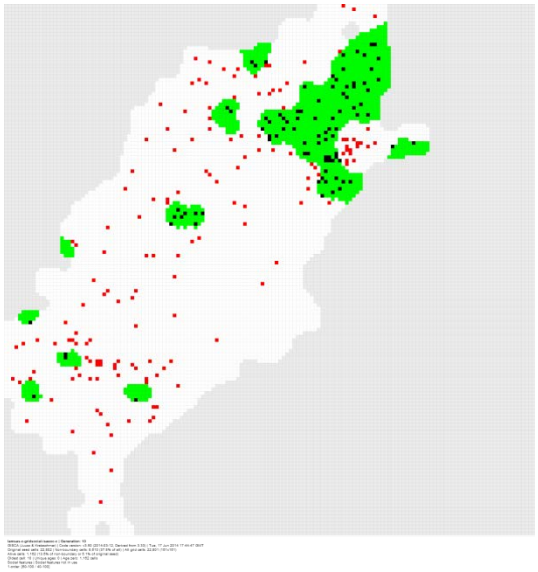

Figure 10 Change in location status for *thundershower*, 10 iterations

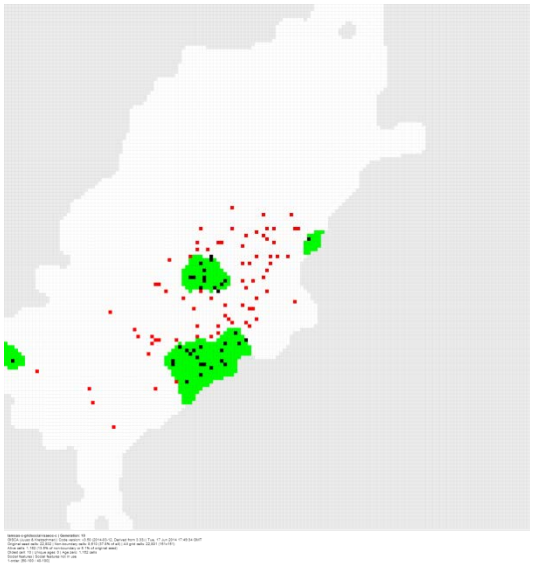

Figure 11 Change in location status for *thundercloud*, 10 iterations

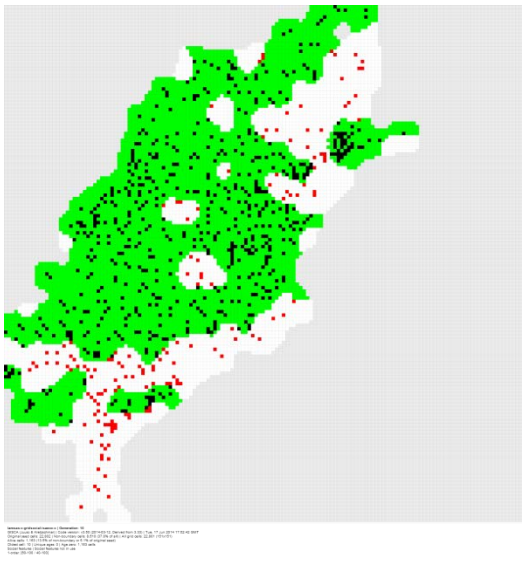

Figure 12 Change in location status for *thunderstorm*, 10 iterations

So far, we have just shown patterns that arise from the thunderstorm item. Figures 13, 14, and 15 show the patterns that the CA creates from three more questions using the R6040-N2 rule set (in color online, in grayscale in the print version). Figure 13 shows the regions created for the living room item, for which

sixteen colors were used for variants out of the 132 elicited for the question. The top five colored regions include *parlor* (orange), *living room* (blue), *sitting room* (pink), *setting room* (light green), and *front room* (yellow).

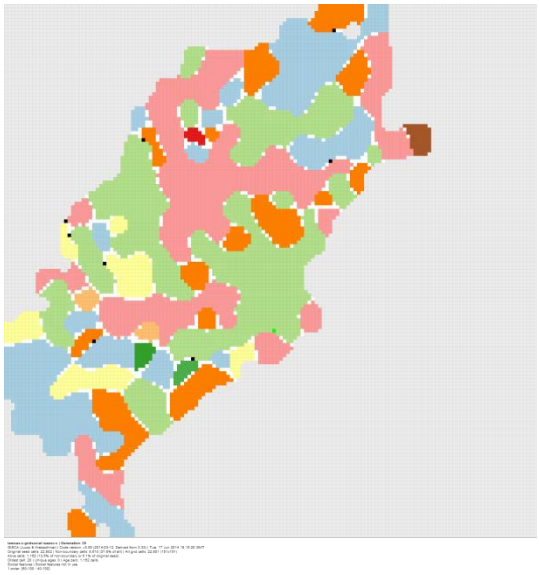

Figure 13 R6040-N2 (20 iterations) for living room question

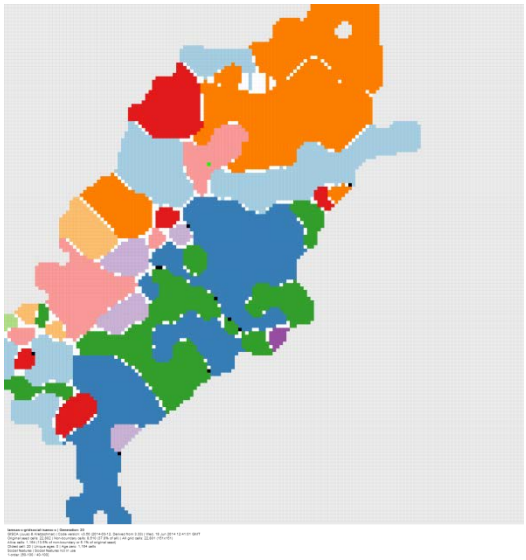

Figure 14 R6040-N2 (20 iterations) for kindling question

Figure 14 shows the patterns created for the kindling item. Eighteeen colors were selected out of the 147 different variants elicited. The top six colored regions included *kindling wood* (orange), *lightwood*

(blue), *rich pine* (pink), *kindling* (light blue), *fat pine* (red), and *lighterd* (green). Both the living room question and the kindling question have many regions, and the same color often occurs in discontinuous regions. These distributions show why the old-fashioned method of using isoglosses is not likely to be very effective as a generalization. Figure 15 shows the patterns created for the sofa question. Unlike the previous items, out of the 121 different response types there is one large region shown for *sofa* (yellow), with a number of smaller regions such as *settee* (orange), *davenport* (red), *couch* (light blue), *bench* (green), *lounge* (lilac), and *loveseat* (brown). All the questions have a great many response types out of which just a few generate regions.

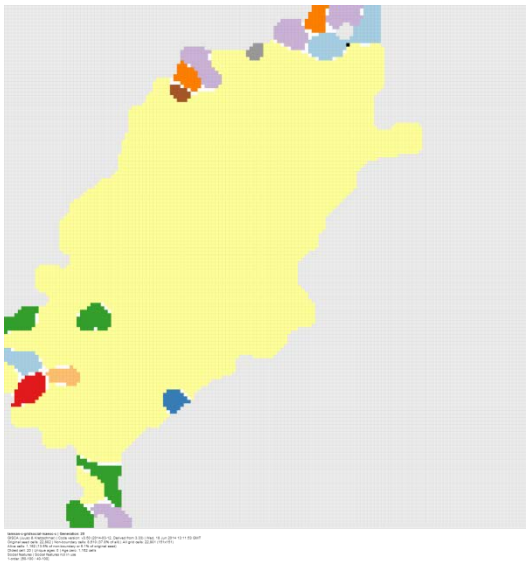

Figure 15 R6040-N2 (20 iterations) for sofa question

In the above examples we have focused on majority vote scenarios, as these can be intuitively understood to best match the call to create generalizations on the original data. The CA does however support all possible rule combinations between 0% and 100% for both adoption and maintenance separately. Lowering the adoption threshold below 50% increases the likelihood of overlap between regions, and decreasing the threshold to 0% will make all variants become adopted at every location across the grid. Similarly, by adjusting the maintenance rule to 0% each variant will be maintained at every location it has existed at any point during the running of the CA, which in turn causes black cells,

i.e. locations with multiple variants, to appear inside larger regions and not just between them. Different rules may prove useful for other purposes than creation of an array of regions.

In the discussion so far we have demonstrated that a CA is practical tool for generating region estimates based on real data and simple task-specific rules. The rules required for running the CA are simple to comprehend as they only pertain to the interactions between a single cell and its neighbors at a given state of the CA. Through the repeated application of these rules the user will find out what the distributional patterns of each known variant will look like under the scenario they defined. Our CA thus provides a mechanism for creating objective and reproducible estimates that are also comparable across other question items in the same survey and other similarly prepared surveys.

4 Experimental evaluation

Considering the range of applications of CAs from any number of spatially-describable problems to random number generation and the Game of Life, it is pertinent that we examine the variant estimates created by the simulation in contrast to other established forms of generalizations of the same data. Obtaining old-fashioned and necessarily subjective isoglosses drawn from the data is neither practical nor helpful, but comparison with previously created density estimation plots involving the same variants provides a sound frame of reference.

Density estimation techniques (DE) make use of the discriminant analysis statistic (commonly found in statistical software packages like SAS and SPSS). In our linguistic application, for any target feature there will be a set of points (the 1162 locations where informants lived) which can be divided into two classes, those where a particular variant response was elicited and those at which it was not. DE estimates the probability that any given coordinate location will belong to either of the classes, given the known values of the points in the survey. We can use DE to make plots that show the density of occurrence of a target feature, show the probability that the feature variant might occur in any part of the survey region, and estimate comprehensively where a variant might be expected to occur in the survey region. Figures 16, 17, and 18 show three variant estimates produced using DE techniques in earlier work (Kretzschmar and

Light 1996, Kretzschmar 1996). Figure 16 shows the dominant variant *thunderstorm* that shows up in yellow in Figure 9 and in green in Figure 12 in the CA estimates above. In the original dataset this variant appears in 759 out of the 1162 locations and is present throughout the survey area with only small gaps in between. Figures 17 and 18 show two further variants: *thundershower* (red in Figure 9, Figure 10) and *thundercloud* (light green in Figure 9, Figure 11). The original data locations for these variants were given in Figure 3 and Figure 4. In Figures 16, 17, and 18 the probability of eliciting the variant at a given location is indicated by a grayscale value, where the darker the location the more likely a variant is to be elicited there. In each case, the darker areas of the DE plots correspond to the colored regions of the figures shown above for each variant.

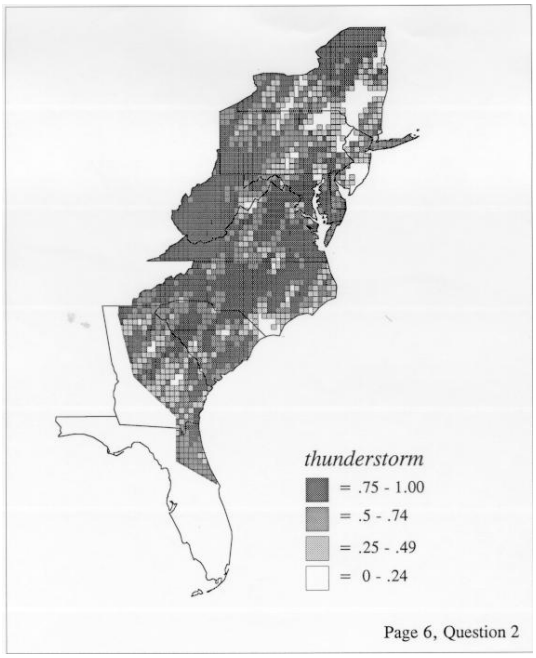

Figure 16 DE plot for *thunderstorm*

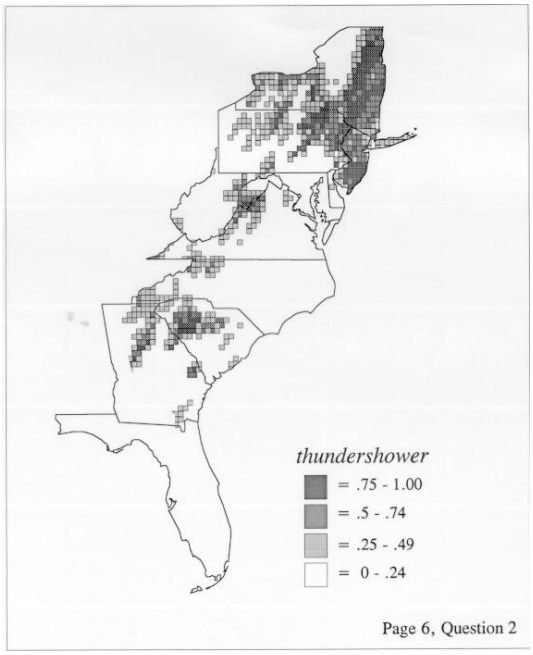

Figure 17 DE plot for *thundershower*

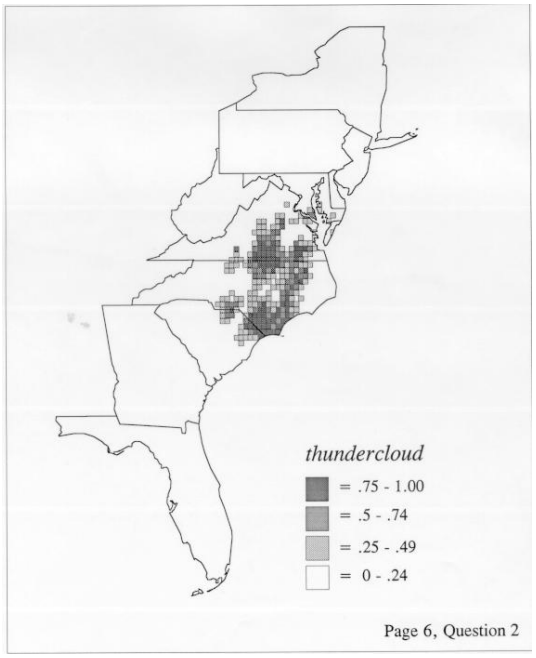

Figure 18 DE plot for *thundercloud*

Figures 19, 20, and 21 show the corresponding variant regions created by the CA using rule R9010-N2, as also shown in Figure 7. Here for clarity the color information is removed and all locations with the given variant show up in black for each figure. The overall map shapes are somewhat different due to some extra padding being introduced in the CA and the Florida peninsula being shown as context in the DE plots, but the survey data used in both is the same. In comparing the two series of images it is apparent that there is strong correlation between the images. Figure 19 shows corridors extending across the darkest patches of the DE plot, scattered locations over some of the medium dark areas, and no locations in the white areas of the DE plot. The same is true for all plot pairs[2], although both Figures 20 and 21 exhibit clusters along the western edge of the survey area that appear to be supported by the original data (as shown in Figure 3 and Figure 4), but do not register in the DE plots. Similarly, some scattered clusters of locations that register as medium grey in the DE plot for *thundershower* do not show up in the CA plot.

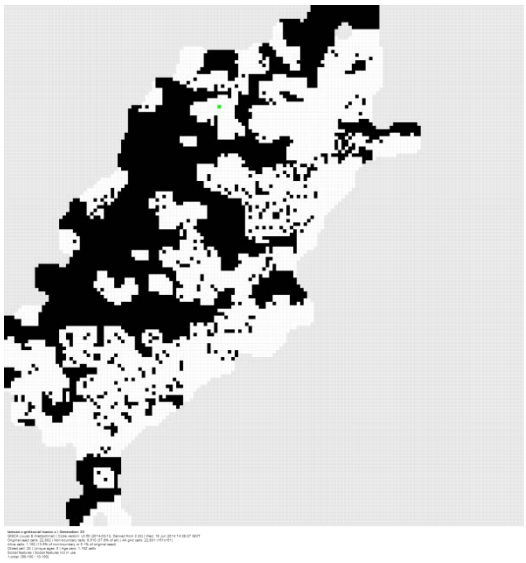

Figure 19 R9010-N2 (20 iterations) for *thunderstorm*

[2] More DE plots are available at http://old.lap.uga.edu/cgi-bin/lapsite.fcgi/lamsas/de-maps/

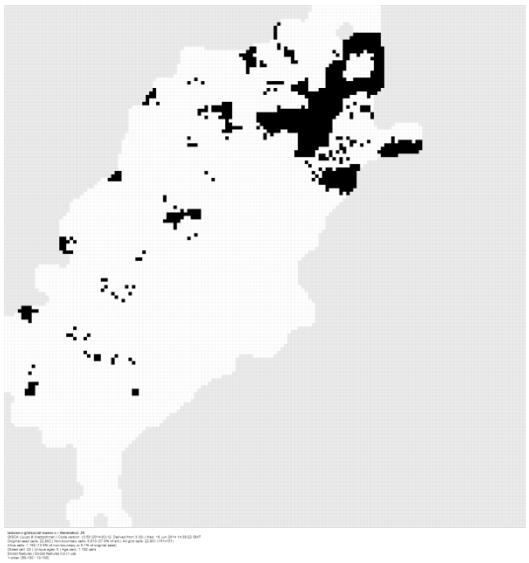

Figure 20 R9010-N2 (20 iterations) for *thundershower*

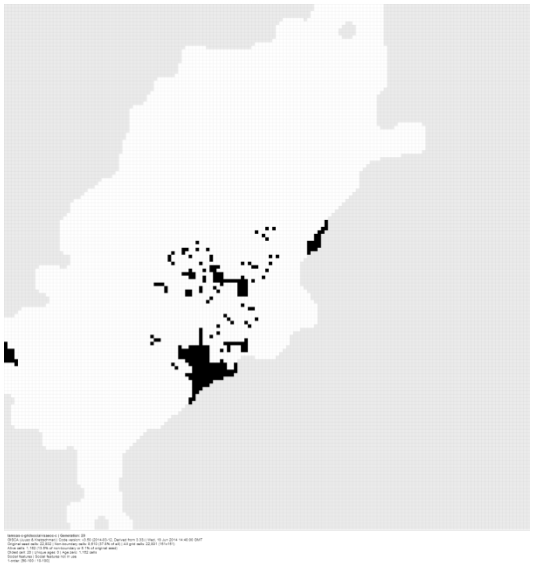

Figure 21 R9010-N2 (20 iterations) for *thundercloud*

It should be noted that while each DE plot was created for a single variant at a time, the CA processes all the variants simultaneously and negotiates regions based on all occurrences of all variants at the same time. The calculation of all 109 individual variant estimates and the resulting overall color region estimate took just seconds to setup, and the required 20 iterations to reach a fixed state were completed in under two minutes on an average laptop computer. The resulting CA plots do not show the density of the

original locations for a variant, but instead the end result of the chosen neighborhood influence scenario in the form of the variant regions that are ultimately negotiated. For the CA, elicitation probability is built into the chosen scenario, which makes the region estimates created by the CA comparable but not equivalent to the probability ranges of the DE plots.

The comparison of the CA region estimates with the earlier DE plots is encouraging. The CA method produces objective generalizations that appear not to disagree substantially with statistical evaluations. The regional patterns created by the CA produce generalized areas of the type that Wenker, Kurath, and others have wanted to produce, only by a completely objective, reproducible method. The exact generalizations created depend heavily on the chosen neighborhood influence scenario, but the scenarios are simple to describe and reproduce without extensive training, which allows for an increased measure of standardization across different data sets and different investigations.

5 Conclusion

Through comparison of corresponding CA and DE plots we have shown that the CA is a practical apparatus for creating region estimates based on real data. The CA differs significantly from the traditional process of drawing isoglosses by providing a rigorous, repeatable process that grows regions based on initial data positions instead of drawing divisions between them as has been the goal in negotiating isoglosses across maps. It would be possible to create further comparisons with the MDS plots mentioned earlier as they too utilize a part of the survey data used here. Similarly, comparison with heat map visualization would be interesting from a data visualization point of view, but direct comparison with the Katz heat maps is not possible due to the differences in the underlying survey data.

It is important to consider that the generalizations presented here respond primarily to the expectations of some linguists and most members of the public that dialect features should be bound to areas, and that one goal of linguistic analysis is to identify those areas. We do not argue that our generalizations represent the truth about variation in language, just that they offer a practical way of identifying regions for dialect features under selected parameters, i.e. the neighborhood influence scenario. Our CA

visualizations create well founded generalizations that answer the question, "Where do people say X?". Generalizations by nature do not show the full complexity of the data, let a alone the full picture of language in the wild, but within these obvious limitations they do provide practical overviews adding to our knowledge of both the data and variation in language. Our CA offers a sound method of producing visualizations to answer one of the questions most often asked of dialect data, in a way that does not require advanced statistical processing and is readily reproducible in an objective and reliable manner.

The work presented in this paper represents one of two major strands of our work on cellular automata. This strand is focused on understanding the real survey data we have loaded into the CA and on the visualization of selected key aspects of it through the apparatus provided by the CA. In addition to the estimation of variant regions presented here, we also use the CA, for example, to visualize the distribution of social features and the impact social similarity between speakers can have in language interaction, again based on real metadata obtained from the survey data. We feel that this data analysis on the original survey data forms a solid foundation for using the same data to create continuously running simulations of language interaction. These simulations are rooted in the theory of complex systems, which forms the other strand of our work on CAs. This second strand attempts to analyze how adaptation and change occur in language over time. Unlike the regions presented here that respond to the popular perception that language is bound to areas, the simulations of our second strand do not quickly reach a fixed state, but rather model the continuous interaction of speakers at many locations as they adopt or discard different dialect features over time. We are pleased to offer a new method for the creation of dialect regions that respond to perceptions of language, and we are also excited by the possibility of simulating not just perceptions but the underlying interactions that form the basis for popular perceptions.

References

Evans, M. R., N. Rajewsky, E. R. Speer. 1999. Exact solution of a cellular automaton for traffic. *Journal of Statistical Physics* 95:1-2, 45-96.

Gardner, Martin. 1970. Mathematical games: The fantastic combinations of John Conway's new solitaire game "Life." *Scientific American* 223: 120–123.

Grieve, Jack, Dirk Speelman and Dirk Geeraerts. 2010. The Emergence of Regional Linguistic Variation: A Computer Simulation. Paper presented at NWAV 39, San Antonio.

Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2013. A Multivariate Spatial Analysis of Vowel Formants in American English. *Journal of Linguistic Geography* 1: 31-51.

Holland, John. 1998. *Emergence: From Chaos to Order*. New York: Basic Books.

Katz, Joshua. 2013. Beyond "Soda, Pop, or Coke:" Regional Dialect Variation in the Continental US. Poster viewed at http://www4.ncsu.edu/~jakatz2/files/dialectposter.png

Keller, Rudi. 1994. *On Language Change: The Invisible Hand in Language*. Trans by Brigitte Nerlich. London: Routledge.

Kretzschmar, William A., Jr. 1992a. Isoglosses and Predictive Modeling. *American Speech* 67: 227-49.

Kretzschmar, William A., Jr. 1992b. Interactive computer mapping for the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS). In N. Doane, J. Hall, and R. Ringler, eds, Old English and New: Essays in Language and Linguistics in Honor of Frederic G. Cassidy (New York: Garland), 400-414.

Kretzschmar, William A., Jr. 1996. Quantitative Areal Analysis of Dialect Features. *Language Variation and Change* 8: 13-39.

Kretzschmar, William A., Jr. 2002. Dialectology and the History of the English Language. In *Studies in the History of English: A Millennial Perspective*, ed. by Donka Minkova and Robert Stockwell (Berlin: Mouton de Gruyter), 79-108.

Kretzschmar, William A., Jr. 2003. Mapping Southern English. *American Speech* 78: 130-149.

Kretzschmar, William A., Jr. 2008. Neural Networks and the Linguistics of Speech. *Interdisciplinary Science Reviews* 33: 336-356.

Kretzschmar, William A., Jr. 2009. *The Linguistics of Speech*. Cambridge: Cambridge University Press.

Kretzschmar, William A., Jr. 2013.Computer Mapping of Language Data. In Manfred Krug and Julia Schlueter, eds., *Research Methods in Language Variation and Change* (Cambridge: Cambridge University Press, 2013), 53-68.

Kretzschmar, William A., Jr., Ilkka Juuso, and C. Thomas Bailey. IP.Computer Simulation of Dialect Feature Diffusion. *Journal of Linguistic Geography*.

Kretzschmar, William A., Jr., and Deanna Light. 1996. Mapping with Numbers. *Journal of English Linguistics* 24: 343-57.

Labov, William, Charles Boberg, and Sherry Ash. 2006. *Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.

Lee, Jay, and William A. Kretzschmar, Jr. 1993. Spatial Analysis of Linguistic Data with GIS Functions. *International Journal of Geographical Information Systems* 7: 541-560.

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten and Willem van de Vis. 1996. Phonetic Distance between Dutch Dialects. In G.Durieux, W.Daelemans, and S.Gillis, eds., *CLIN VI: Proc. of the Sixth CLIN Meeting* (Antwerp: Centre for Dutch Language and Speech (UIA), 185-202.

Nerbonne, John, and Peter Kleiweg. 2003. Lexical Distance in LAMSAS. In John Nerbonne and William A. Kretzschmar, Jr., eds., *Computational Methods in Dialectometry*, special issue of *Computers and the Humanities* 37(3): 339-357.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg and Therese Leinonen. 2011.

Gabmap — A Web Application for Dialectology. *Dialectologia.* Special Issue II: 65-89.

Sarkar, Palash. 2000. A Brief History of Cellular Automata. *ACM Computing Surveys* 32: 80-107.

Schneider, Edgar, 1988. "Qualitative vs. quantitative methods of area delimitation in dialectology: a comparison based on lexical data from Georgia and Alabama." *Journal of English Linguistics* 21: 175-212.

Thill, J., W. Kretzschmar, I. Casas, and X. Yao. 2008. Detecting Geographic Associations in English Dialect Features in North America with Self-Organising Maps. In *Self-Organising Maps: Applications in GI Science*, ed. by P. Agarwal and A. Skupin (London: Wiley), 87-106.

Tomassini, M., M. Sipper, and M. Perrenoud. 2000. On the generation of high-quality random numbers by two-dimensional cellular automata. *IEEE Transactions on Computers* 49 (10): 1146–1151.

Topa, Pawell. 2006. A Cellular automata approach for modelling complex river systems. In S. El Yacoubi, B. Chopard, and S. Bandini, eds., *ACRWE 2006*, LNCS 4173, pp. 482–491. Berlin: Springer-Verlag.

Torrens, Paul M., and Itzhak Benenson. 2005. Geographic Automata Systems. *International Journal of Geographical Information Science* 19.4: 385-412.

Von Neumann, J. and A. W. Burks. 1966. *Theory of Self-Reproducing Automata*. Urbana: University of Illinois Press.

Wolfram, Stephen. 2002. *A New Kind of Science*. Champaign, IL: Wolfram Media.