Local Spatial Biclustering and Prediction of Urban Juvenile Delinquency and Recidivism

Alan J. Izenman[1*], Philip W. Harris[2], Jeremy Mennis[3], Joseph Jupin[4] and Zoran Obradovic[4]

[1]*Department of Statistics, Temple University, Philadelphia, PA 19122, USA*

[2]*Department of Criminal Justice, Temple University, Philadelphia, PA 19122, USA*

[3]*Department of Geography and Urban Studies, Temple University, Philadelphia, PA 19122, USA*

[4]*Center for Information Science and Technology, Temple University, Philadelphia, PA 19122, USA*

Received 6 October 2010; revised 21 March 2011; accepted 25 March 2011
DOI:10.1002/sam.10123
Published online 26 April 2011 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Using a novel database, ProDES, developed by the Crime and Justice Research Center at Temple University, this article investigates the relationship between spatial characteristics and juvenile delinquency and recidivism—the proportion of delinquents who commit crimes following completion of a court-ordered program—in Philadelphia, PA. ProDES was originally a case-based sample, where the cases were adjudicated in family court, 1994–2004. For our analysis, we focused attention on studying 6768 juvenile males from the data set. To address the difficult issue of nonstationarity in the data, we considered various two-way clustering algorithms to group the juveniles into 'types' by way of the many variables that described the juveniles. Following different modeling scenarios, we applied the plaid biclustering algorithm in which a sequence of subsets ('layers') of both juveniles and variables are extracted from the data one layer at a time, but where overlapping layers are allowed. This type of 'biclustering' is a new way of studying juvenile-offense data. We show that the juveniles within each layer can be viewed as spatially clustered. The layers were determined as descriptive tools to aid in identifying subsets of the data that could be useful in policy making. Statistical relationships of the variables and juveniles within each layer are then studied using neural network models. Results indicate that the methods of this paper are more successful in predicting juvenile recidivism in urban environments when different crimes are modeled as separate data sets rather than being pooled together as a single data set. © 2011 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 4: 259–275, 2011

Keywords: biclustering; data mining; Getis–Ord statistic; neural networks; nonstationarity; plaid models; ProDES database; spatial statistics; two-way clustering; urban crime

1. INTRODUCTION

One of the most challenging aspects of analyzing social science data is the presence of spatial effects in quantitative models of individual behavior. Recent research [1–4] has shown that one's surroundings can play a key role in determining individual behavioral outcomes in a variety of contexts, such as health and crime. Unfortunately, quantitative research in the social sciences that focuses on studying spatial effects has been quite limited in scope, due primarily to the particular challenges associated with analyzing spatial data. Such challenges include issues of data quality, data integration, and the use of appropriate statistical techniques. In addition, the large size, high dimensionality, and complexity of many current social science data sets are problematic for developing informative and parsimonious models of behavioral outcomes.

Certainly, conventional statistical approaches have been adapted to incorporate spatial effects [5]; for example, in the use of spatial econometrics [6] and hierarchical linear modeling [7]. However, such approaches typically either treat spatial effects as a nuisance to be controlled (so as to obtain an unbiased conventional model) or are subject to assumptions regarding the nature of the spatial relationships in question that may indeed mask relevant neighborhood influences on individual outcomes [8]. Additionally, these

Correspondence to: Alan J. Izenman (alan@temple.edu)

adaptations of conventional modeling approaches do not address the analytical challenges associated with very large, high-dimensional, and noisy data sets.

In this article, we investigate the interplay between spatial and individual effects in the prediction of juvenile delinquency and recidivism. Understanding how individual and spatial characteristics shape youth behavior is fundamental to planning programs that facilitate positive trajectories for physical, social, cognitive, and affective youth development. Our results indicate that certain groups of juveniles are particularly susceptible to specific causal mechanisms of recidivism that should be considered by the court at time of 'disposition' (the juvenile equivalent of sentencing). For example, removal from the community may be beneficial to juveniles in a group in which peer influence is seen to be a major factor in causing recidivism. On the other hand, if a juvenile is a member of a group defined by, say, parental criminality or substance abuse, other rehabilitation approaches should be taken to address those mechanisms. As far as we know, the courts do not consider these differences in any coherent way. In other words, this research is intended as basic research, but it has implications for how a court may sentence juveniles more effectively based upon the likely mechanisms of recidivism for different situations.

The remainder of this article is organized as follows. Section 2 gives some background to the current study. The Program Development and Evaluation System (ProDES) database, which contains information on all cases of juvenile delinquency and recidivism in Philadelphia during the period 1994 and 2004, is described in Section 3. The greatest part of this research effort, as is common with data-mining studies, was concerned with the preparation of the data set for use in this study; specifically, the selection of variables, preprocessing the data, data reduction, and a fundamental refocusing of the objectives of the study. Nonstationarity in the juvenile recidivism data emerged as a central issue and was accounted for by fitting a plaid biclustering model; the results are described in Section 4. Neural network modeling of each plaid layer and estimates of prediction error are detailed in Section 5. An assessment of the accuracy of the modeling process is given in Section 6, and a concluding discussion is given in Section 7. An outline of the entire data-analysis process is displayed in the flowchart in Fig. 1.

2. BACKGROUND

One of the most difficult challenges facing researchers on the role of spatial effects in a variety of social science domains is being able to integrate individual and spatial information to encode spatial relationships. Administrative records of an individual's characteristics often contain a georeference (or locational coordinate) for that individual, such as an address. In order to retrieve spatial characteristics for such individuals, the administrative records must be matched to other spatial information, such as socioeconomic data from the US Bureau of the Census or criminal-activity data from a municipal police department. We focus our attention on investigating the combined effects of home and local environments and individual characteristics on continued delinquent behavior. Estimating the effect of

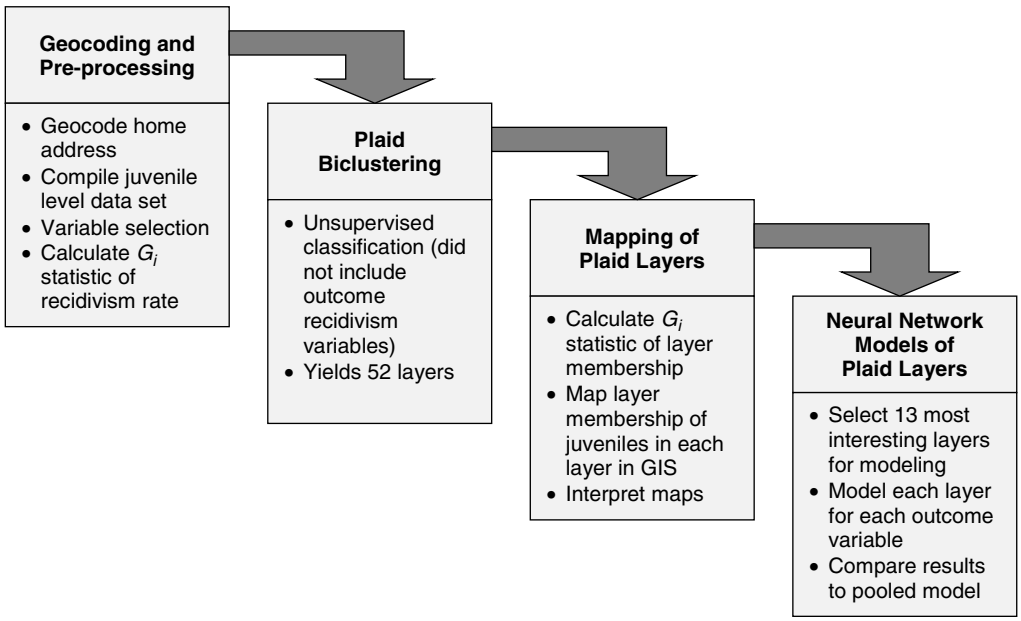

Fig. 1 Flow chart of the data-analysis process.

the local environment on the likelihood of recidivism has become increasingly important because 'aftercare services' have proliferated for adjudicated youth. At present, the analytical tools that researchers have been using are inadequate for this type of research.

Two facts guide our thinking: first, adolescent development and behavior can be supported or hampered by environmental forces [9–11] and second, delinquent youths are over-represented in neighborhoods characterized by disorganization or criminogenic organization [1]. To the extent that environmental forces impede social, emotional, and physical development, programs for delinquent youths can intervene to increase individual and social efficacy; these programs also serve as a buffer between youths and harmful external forces, so that natural developmental processes can continue. This view of intervention programs is particularly important in light of the finding that in neighborhoods characterized by poverty and social disorganization, residents are less willing to intervene when they see youths engaging in antisocial or unlawful acts [12]. In addition to these environmental effects, recent research has demonstrated the potential deleterious effect of the actual institutional placements on child development [13]. This implies that aftercare services must address the youths' developmental needs, which may be aggravated by a period of institutionalization, as well as the external forces that inevitably compete with program effects.

Wilson's 1987 book *The Truly Disadvantaged* [14] stimulated a flurry of academic activity examining the role of the local environment in producing a host of outcomes, including educational attainment, cognitive skills, early or unplanned pregnancy or parenting, and labor-market success [11,15–20]. Concurrently, there was a resurgence of interest in social disorganization theory [21], which highlighted the role of the local environment in promoting or prohibiting crime and delinquency through (a lack of) cohesion among neighbors and community-level social control. Despite this fact, correctional scholarship has not examined the role that local environments play in reinforcing or weakening the treatment effects of these interventions. A review of the outcomes for youth aftercare programs suggests that half of their clients re-offend at some time during the year after release, and one-third return during this time to a more secure placement.

3. THE DATA SET

3.1. The ProDES Database

The ProDES database is a population database of all juvenile cases committed by the Philadelphia Family Court to community and residential programs between 1994 and 2004. ProDES was a project of the Crime and Justice Research Center (CJRC), Temple University, funded by the Department of Human Services during 1994–2004. The ProDES database tracked juveniles assigned to court-ordered programs by the Family Court of Philadelphia, PA, and was designed to evaluate all programs used by the City of Philadelphia for its delinquent youth. 'Delinquent' is a status that includes delinquent acts or offenses as well as a judgment that the youth requires supervision beyond that being provided by the parents. ProDES was designed to provide outcome information to programs for delinquent youths and to users of these programs, namely judges, probation officers and funding agents. Its goals are to provide continual feedback to the programs and to the juvenile court that will facilitate program development, facilitate better matching of youths to programs, and identify and facilitate improvements in the array of programs available to the Philadelphia juvenile-justice system.

Following a youth's arrest, charging, detention, pretrial hearing, and a trial (in which the judge not only must find the youth guilty but must also find the youth to be a delinquent), the following options are available to a judge at the disposition of a case: (i) Probation: the youth will continue to live at home or in the home of a relative, and is supervised by a probation officer; (ii) Foster care: the youth is placed in an approved foster home for a period of time; (iii) Community-based program: the youth is required by the court to attend a program (after-school program, an alternative program, or a mentoring program) in the community; (iv) Residential facility: the youth is removed from his/her home and placed in a residential facility with other delinquent youths; (v) Aftercare: after completing a required period of time in a residential facility, the youth returns to court for a second disposition on the same offense and is committed to an aftercare program designed for youths reentering the community following a period of being incarcerated. Of primary interest in our study are options (iii) and (v), because these are the cases in which the youth is in a program while living at home.

ProDES collected data at four points in time: (i) at the point of disposition (the juvenile equivalent of sentencing), data are extracted from the youth's record that contains information such as offense history, placement history, needs (e.g., drug use, mental health problems), and family history; (ii) at program intake, staff persons are asked to complete a needs assessment and the youth completes a self-report section containing psychometric scales; (iv) at discharge, the intake process is repeated and program staff report on the youth's progress in the program; and (iv) 6 months following program discharge, a follow-up record check is conducted to identify any new petitions (arrests leading to charges) generated in the juvenile or adult court systems, and telephone interviews are conducted with youths, when available, and guardians. Although

the juveniles in our data set range in age from 10 to 20 years old, the majority (69%) are between 15 and 17 years old, predominantly male (90%) and African-American (73%). The data include measures of family demographics, juvenile characteristics, criminal history, current offense characteristics, recidivism status, and many other items. The program intake and program discharge data were collected by program staff who were trained by staff of the CJRC, using instruments developed by CJRC. All other data were collected by CJRC staff. Variables that identified the juvenile subjects were removed from the database for use by the study researchers. The cases in ProDES were geocoded (using ArcView GIS 9.2) based on the home address (and zip code) given at the point of disposition listed for the juvenile. The success rate for geocoding was 98% after manually addressing errors such as misspelled street names. We also restricted our analysis to cases that had been in the system for at least 6 months, so as to examine only those cases that had the possibility of recidivating. More information on this project can be found at http://www.temple.edu/prodes.

Initial research interest focused on a collection of 45 585 cases and about 1200 variables. Although this study began with the intent of a case-based analysis of the data, the entire database proved to be too unwieldy for modeling or predicting recidivism accurately because cases were not distinct juveniles. Accordingly, we decided to study a related question using a juvenile-based approach. A total of 13 418 juvenile records were selected from the period between 1996 and 2002—the years when the data were most complete. The data set was further reduced by the removal of cases involving females, as prior research [22–24] (reinforced by our own analysis) demonstrated a gender difference concerning the predictors of juvenile delinquency and recidivism. We selected the first-occurring case for each juvenile and deleted juveniles with incomplete records. These considerations resulted in a sample of size 6768 drawn from the all-male juvenile population who had been remanded to programs within their communities by the Philadelphia Family Court.

3.2. Outcome Variables

We consider several measures of juvenile delinquency. The primary such measure is *any type of recidivism* (coded as 'ganypet'); specifically, delinquent juveniles who commit any type of violation while in a court-ordered, community-based program or within 6 months after completion of that program. These violations can range in severity from a felony criminal offense to a probation violation. The recidivism rate is defined as the ratio of the number of recidivating cases to the total number of delinquent cases within a given area. Offense-specific measures are defined by the recidivating violation. The first is *personal-offense recidivism* (coded as 'xperson'), delinquents who recidivate by committing a crime classified as a personal offense. Personal offenses are violent crimes committed against a person (e.g., robbery or assault) and, thus, indicate cases at particular risk. The second is *drug-offense recidivism* (coded as 'xdrug'), delinquents who recidivate by committing a drug crime. The third is *property-offense recidivism*, delinquents who recidivate by committing a property crime (coded as 'xproperty').

To give some idea of where recidivism occurs within Philadelphia, Fig. 2 shows an annotated map of its 45 nonoverlapping neighborhoods superimposed upon the residences of the juvenile delinquents in the data set. The roots of identity for many of the inner-city neighborhoods date from the 19th and even 18th centuries, while other neighborhoods represent more recent 1950s and 1960s housing developments. The boundaries of these neighborhoods are typically major natural and human-made features, such as rivers and major roads and highways. One can assume a relatively high level of within-neighborhood demographic homogeneity, as most urban neighborhoods contain residents of similar race and class. This is particularly true in Philadelphia, which remains highly segregated by race and class, with certain exceptions. High recidivism rates cluster in Kensington, Richmond, and Hunting Park, as well as in Wynnefield and around Pennsport. Low recidivism rates occur mostly along the far northern tier of the city from Chestnut Hill through Oak Lane to the far northeast region of the city.

4. NONSTATIONARITY AND CLUSTERING: ASSESSING TYPES OF JUVENILE DELINQUENTS

Classification of persons or cases is central to studies of behavior. Reasons for classifying juvenile delinquents include improving our understanding of delinquent behavior, matching offenders to interventions, managing offender populations, and improving risk prediction. An undifferentiated examination of delinquency patterns or of individual delinquents is likely to mask relevant information about who is positively or negatively affected by what. Differences among offenders and their individual circumstances will affect responses to specific intervention methods. Knowing what works and what does not work with different types of individuals under different circumstances continues to be the most critical goal of program evaluation in juvenile corrections and delinquency prevention.

4.1. Input Variables

Most of the original predictor variables were dropped as input to a cluster analysis based upon considerations from alternative analyses. The variables used were chosen based

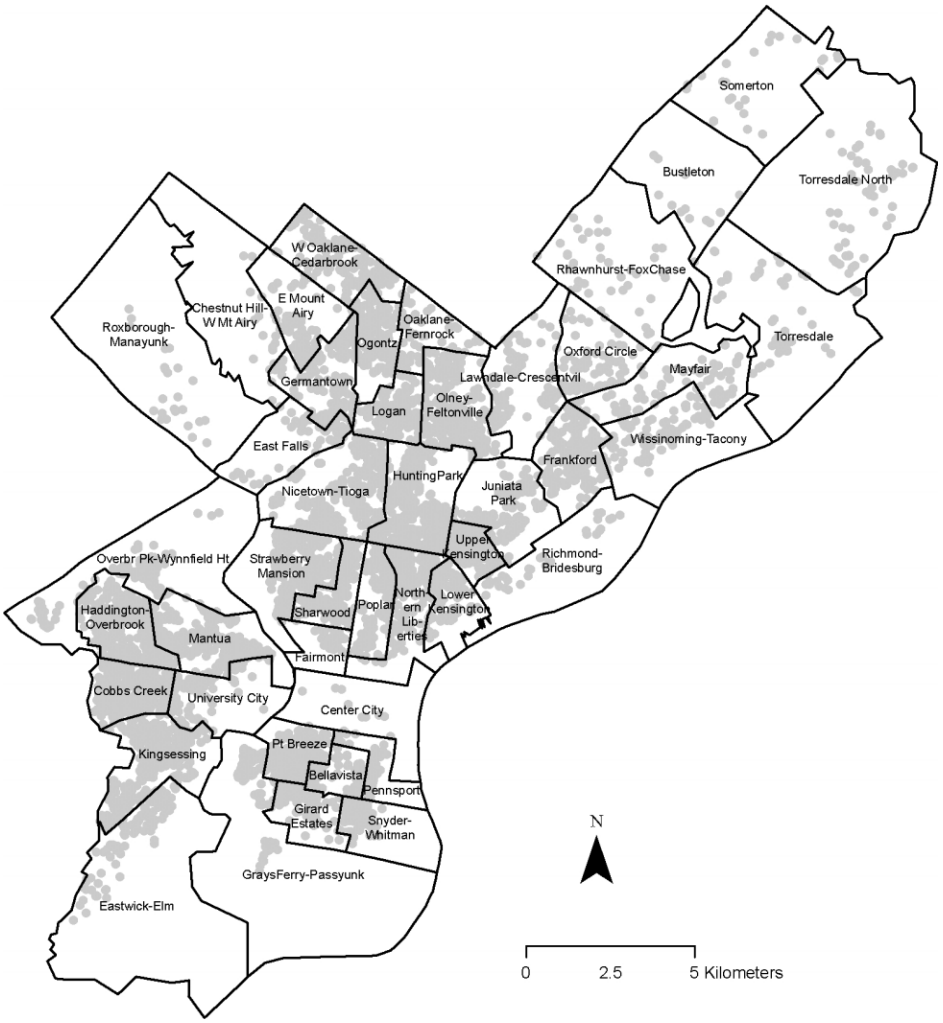

Fig. 2 Map of Philadelphia neighborhoods. Gray dots indicate the residences of juvenile delinquents in the data set.

upon the juvenile-justice literature, recommendations of domain experts, and several stages of variable selection and modeling, including the use of logistic regression, decision trees, and neural networks [25]. The 27 variables used as input to a cluster analysis are listed in Table 1. The input variables were of four types: (i) background characteristics of the individual juvenile; (ii) the initial offense that the juvenile committed upon entry to the Family Court system (referred to as the 'instant offense'); (iii) indicators of social disorganization within the neighborhood within which the juvenile resides; and (iv) indicators of overall delinquency and recidivism nearby the juvenile's home (referred to as 'contagion' variables). Nineteen of the 27 variables were descriptors of the juvenile and eight were spatial descriptors of the local home environment of the juvenile. Variables describing background characteristics of the individual juvenile include basic descriptors of age and race. The juvenile's family history regarding crime was captured by a variable indicating whether a parent of the juvenile had a criminal record. The juvenile's own delinquency history was captured using variables that indicated the number of prior arrests (note that a juvenile may have been previously arrested but not sent to a court-ordered program), whether the juvenile was living in an institution (as opposed to with his family or other living arrangement) immediately prior to the targeted community-based case, and whether the juvenile had any prior out-of-home placement. Note that the 'lives in an institution' variable indicates a juvenile with severe-enough delinquent behavior or other issues for a judge to decide that it is in the community's best interests to remove the juvenile from his home.

The instant offense for each juvenile was coded in the same manner as the outcome variables. Social disorganization of the juvenile's residential neighborhood was captured using crime, housing, and socioeconomic data. Block-level addresses of arrest data for the period 2000–2002 were

Table 1. The input variables used in the plaid analysis of juvenile recidivism.

Variable	Description
Individual characteristics	
WhiteDum	Is youth white?
HispanicDum	Is youth Hispanic?
Probation	Was youth on probation at the time of his arrest?
LiveInstitution	Did the youth live in an institution?
PriorPersonalChgs	Did the youth have prior personal offense charges?
Juvdrgar	Did the youth have prior drug arrests?
Prioroutofhomepl	Was the youth placed in out-of-home program?
sibarr	Did any of the youth's siblings have an arrest record?
jhismh	Did the youth have a history of mental-health problems?
age	How old was the youth at the case recording?
AlchoholAbuse	Did the youth have a history of alcohol-abuse?
DrugAbuse	Did the youth have a history of drug-abuse?
Family history	
ParenDeceased	Is at least one parent deceased?
ParSubAbuse	Did the parents have a history of substance-abuse?
ParentalCrime	Does a parent have a criminal history?
Instant offense	
sexoff	Was the instant offense a sexual offence?
InstantPerson	Was the instant offense person-related?
InstantProperty	Was the instant offense property-related?
victinj	Did the youth injure a victim in the instant offense?
Social disorganization	
den_dr_sale	Density of drug arrests within 500 m of youth's home
den_person	Density of person offenses within 500 m of youth's home
Local environment	
p_black	Percent black in census block
p_vacant	Percent of vacant housing in census block
p_spanish	Percent Hispanic in census block
p_highsch	Percent high-school graduates in census block
Spatial contagion	
kcnt_1 km	Number of delinquent youths residing within 1 km of the juvenile
G_i (recidivism clustering)	z-value of Getis-Ord Gi statistic applied to the neighborhood recidivism rate

acquired from the Philadelphia Police Department as text addresses and geocoded. We then calculated for each juvenile the density of arrests in their home neighborhood by summing the number of arrests within 500 m of each juvenile's home and dividing that number by the area of the circular neighborhood defined by that 500 m radius. We focused on two types of police-data arrests, namely, drug-sale arrests and personal-offense arrests, because both variables are indicative of social disorganization. We also included three housing variables (percentage African Americans, percentage Hispanic, percentage of vacant housing units) and a socioeconomic variable (the percentage over the age of 25 with a high-school diploma or equivalent) derived from US Bureau of the Census 2000 block-level data. These variables are intended to reflect family organization, housing infrastructure, and educational attainment at the neighborhood level. We also considered that the likelihood of a juvenile recidivating may be influenced not only by his own characteristics, but also by the behavior of juveniles living nearby. We generated two 'spatial contagion' variables designed to capture this effect: the number of juvenile delinquents who live within 1 km of the juvenile's home and the z-score of the Getis–Ord G_i statistic [26,27] applied to the juvenile recidivism data.

Define $\mathcal{N}_i$ to be a circle of radius $d = 1$ km centered at the ith juvenile's home and set X_i (with value x_i) to be the ratio of the number of recidivating juveniles to the total number of juveniles within $\mathcal{N}_i$, $i = 1, 2, \ldots, n$, where $n = 6768$. Define the $(n \times n)$ symmetric weight matrix $\mathbf{W} = (w_{ij})$ by $w_{ij} = 1$ if the jth juvenile is in $\mathcal{N}_i$, and 0 otherwise. The Getis–Ord G_i statistic is defined as

$$G_i = \frac{\sum_{j=1, j\neq i}^{n} w_{ij} X_j}{\sum_{j=1, j\neq i}^{n} X_j}, \quad i = 1, 2, \ldots, n, \tag{1}$$

so that $G_i \in [0, 1]$ is a measure of local spatial autocorrelation that indicates if the values within the spatial neighborhood around the ith juvenile differ significantly from the data set as a whole. Values of G_i close to 1 indicate a clustering of high values, while G_i values close to 0 indicate a clustering of low values. Note that G_i does not include the ith juvenile. It was shown by Getis

and Ord [26] using a permutation approach that, under spatial independence, holding $\overline{x}_i = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} x_j$, and $s_i^2 = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}(x_j - \overline{x}_i)^2$ fixed, the mean and variance of the permutation distribution of G_i are, respectively,

$$E(G_i) = \frac{w_i}{n-1}, \quad \text{var}(G_i) = \frac{w_i(n-1-w_i)}{(n-1)^2(n-2)}\left(\frac{s_i}{\overline{x}_i}\right)^2, \tag{2}$$

where $w_i = \sum_{j=1,j\neq i}^{n} w_{ij}$. When there is no spatial clustering among the X_i, Zhang [28] showed that the permutation distribution of $Z_i = (G_i - E(G_i))/\sqrt{var(G_i)}$ is approximately Gaussian for large n. (Gaussianity can fail, however, if d is taken to be either too small or too large, or if the underlying distribution is substantially skewed.) The statistic Z_i is used here to measure the degree of local spatial clustering of the recidivism rate. A value of $Z_i > 2.0$ indicates a high degree of local spatial clustering of high recidivism rate (referred to as a 'hotspot'), while a value of $Z_i < -2.0$ indicates a high degree of spatial clustering of low recidivism rate (a 'coldspot'). For an excellent discussion of hotspots in mapping crime, see ref. 29. After trying different scaling strategies, all input variables were scaled before further analysis to have their smallest value -1 and largest value $+1$.

4.2. Nonstationarity

It has long been understood [30] that spatial data such as we have for our study typically do not conform to the usual nonspatial assumptions of independence, homogeneity, and stationarity. For such data, neighboring observations are not independent of each other. While homogeneous data randomly scattered over a number of locations tend to exhibit stationarity in their distribution of values, spatial data typically display characteristics of nonstationarity. Indeed, for such data, we often see high values tend to cluster together just as low values cluster together, and such similarity between observations tends to dissipate as the observations become more distant geographically from one another. We expect relationships between variables to differ from location to location, so that a global model would likely provide inaccurate estimates of relationship strengths while also failing to account for important local patterns. Prediction for our data set is difficult because there are subgroups in the data for which particular relationships hold between the variables. In other words, relationships between the variables differ depending upon the subgroups. Some of this 'nonstationarity' is spatial in nature because it is related to race and class and other characteristics with strong spatial dependency [31]. But some of the nonstationarity may concern characteristics that are not spatially dependent, such as parent substance abuse. In our study, all available evidence points to nonstationarity of recidivism patterns of juveniles, so that taking nonstationarity into account should improve predictive power. Our approach first addresses the nonstationarity issue by clustering the juveniles into different 'types' (identified by specific subsets of the variables) and then developing risk prediction within each type.

4.3. Two-Way Clustering

Classical clustering techniques try to divide up all the sample individuals into nonoverlapping, homogeneous groups based upon information on those individuals provided by a given set of variables; this objective is usually accomplished by applying one (or more) of a large collection of one-way clustering algorithms to the individuals in the data; see, for example, Ch. 12 of ref. 32. However, for some situations, such a one-way partition of the data may not be the most appropriate technique to apply to the data. What may be more relevant is to perform a two-way clustering of the individuals and the variables. This strategy may be carried out by a one-way clustering, first, on the rows of the data (the individuals), followed by a one-way clustering of the columns (the variables), and then try to reconcile the results. This strategy may not yield useful findings if the rows and columns are dependent upon each other in some unusual and complex way (as in our study).

The development of two-way clustering algorithms in which rows and columns can be clustered simultaneously has been studied extensively in the statistical and computer science literature, and such algorithms have been dubbed either as 'co-clustering' [33–36] or as 'biclustering' [37–44] algorithms. For a detailed review of these methods, see Section 3.2 of ref. 45. The essential differences between these types of two-way clustering include the following: co-clustering clusters all the rows and all the columns of a data set simultaneously, clusters are nonoverlapping (in the sense that rows or columns cannot be members of more than one cluster), and all rows and columns must be accounted for in the results; biclustering, on the other hand, relaxes the exhaustive nature of that approach by permitting some rows and some columns not to be included in any of the clusters (because they may be redundant or noninformative for the clustering process), and overlapping clusters are allowed. A bicluster has been characterized [44] as a submatrix of the data matrix whose entries satisfy some prespecified condition, whose rows and columns need not be contiguous, where different submatrices may overlap one another, and some rows and columns may be omitted from the selection process. General consensus appears to favor biclustering as a better approach to many scientific problems and, in particular, the 'plaid' algorithm is regarded as one of the most useful ways of discovering biclusters from microarrays and other similarly structured data.

4.4. Plaid Biclustering Models and Algorithm

The 'plaid' biclustering algorithm [40] has been used successfully for two-way biclustering of gene-expression data, nutrition data, financial data, and repeated-measures data, and is recognized as one of the best biclustering ideas. See also Section 12.8.2 of ref. 32. Some follow-up work has started to appear on plaid models; see Turner [45], Turner *et al*. [46,47], and, with a Bayesian version of plaid, Caldas and Kaski [48]. Shabalin *et al*. [44] studied a large number of biclustering algorithms and compared the performances of those algorithms when applied to both real (gene expression levels) and simulated data; they found *inter alia* that Plaid performed very well and was one of only two such algorithms that could handle large quantities of data and generalize to higher-dimensional data arrays. So far, all the published examples illustrating plaid models have used only continuous variables. In our case, as with most social-science data, the variables constitute a mixture of binary-valued and continuous measurements. Fortunately, the plaid algorithm makes no assumption (explicit or implicit) that the variables have to be continuous or have to be generated from a particular probability distribution. (Compare this with the fact that certain biclustering algorithms [49] are specifically constructed to be applied to gene expression levels, which are assumed to follow a Gaussian distribution.) As we will see, the plaid algorithm is an iterative version of least-squares, where a quadratic error function Q is minimized using only calculus and without any reference to an underlying distribution. Only if we were interested in significance testing of the model parameters (which we are not) would we need to make distributional assumptions.

The plaid model partitions the data into a sequence of biclusters or 'layers'. Each layer is formed from a subset of the rows and a subset of the columns, and can be viewed as a two-way clustering of the elements of the data array, except that rows (individuals) and columns (variables) can be members of different layers or of none of them. Hence, overlapping layers are allowed. Depending upon the data and the algorithm settings, the number of layers can vary quite a bit, from one layer to over 50 layers in some situations. Although approaches to biclustering with overlapping layers have appeared in the scientific literature, this idea appears to be unknown in social-science research.

Let X_{ij} denote the value of the ith juvenile measured on the jth variable. The plaid model can be written approximately as a sum of several terms,

$$X_{ij} \approx \theta_{ij0} + \sum_{k=1}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}, \tag{3}$$

where θ_{ij0} is an overall effect term and the terms in the sum are called 'layers'. The kth term in the sum refers to the kth layer and consists of a weight function θ_{ijk} times the product of two indicator functions, ρ_{ik} and κ_{jk}. The indicator function ρ_{ik}, is equal to 1 if the ith juvenile is in the kth layer, and is zero otherwise. The other indicator function, κ_{jk}, is equal to 1 if the jth variable is in the kth layer, and is zero otherwise. So, a term will only be present in the sum if both indicator functions equal 1; that is, if both the ith juvenile and the jth variable are simultaneously in the kth layer. The weight function for the kth layer can be expressed in a variety of different ways, but here we use the two-way additive representation, $\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk}$, of a layer effect ($\mu_k$) plus a row effect ($\alpha_{ik}$) plus a column effect ($\beta_{jk}$), $k = 0, 1, 2, \ldots, K$, where $k = 0$ is taken to be a 'background' layer. The plaid model can, therefore, be written as

$$X_{ij} \approx (\mu_0 + \alpha_{ij0} + \beta_{jk0}) + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk})\rho_{ik}\kappa_{jk}, \tag{4}$$

where, to avoid overparametrization, we require $\sum_i \rho_{ik}\alpha_{ik} = \sum_j \kappa_{jk}\beta_{jk} = 0$. An error sum-of-squares criterion,

$$Q = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(X_{ij} - \theta_{ij0} - \sum_{k=1}^{K}\theta_{ijk}\rho_{ik}\kappa_{jk}\right)^2 \tag{5}$$

is used to estimate the various unknown plaid model parameters from the data, where each term is the squared error in using the plaid model to predict the observed entry in a particular row and column, summed over all r columns and all n rows. For a large number K of layers, the optimization problem quickly becomes computationally infeasible: each row or column can be in or out of each layer, which means that there are $(2^n - 1)(2^r - 1)$ possible combinations of rows and columns to consider. To resolve this computational problem, the minimization of the criterion Q is accomplished by an alternating least-squares iterative process, in which one layer is estimated at a time.

Suppose we have already fitted $K - 1$ layers, and we need to identify the Kth layer by minimizing Q. If we let $E_{ij} = X_{ij} - \theta_{ij0} - \sum_{k=1}^{K-1}\theta_{ijk}\rho_{ik}\kappa_{jk}$ denote the 'residual' remaining after fitting the first $K - 1$ layers, then we can write Q as

$$Q = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{r}\left(E_{ij} - \theta_{ijK}\rho_{iK}\kappa_{jK}\right)^2 \tag{6}$$

$$= \frac{1}{2}\sum_{i=1}^{r}\sum_{j=1}^{n}\left(E_{ij} - (\mu_K + \alpha_{iK} + \beta_{jK})\rho_{iK}\kappa_{jK}\right)^2. \tag{7}$$

We wish to minimize Q subject to the identifying conditions

$$\sum_{i=1}^{n}\alpha_{iK}\rho_{iK}^2 = \sum_{j=1}^{r}\beta_{jK}\kappa_{jK}^2 = 0. \tag{8}$$

From Eqs. (7) and (8), we set up the usual Lagrangian multipliers, differentiate with respect to μ_K, α_{iK}, and β_{jK}, set the derivatives equal to zero, and solve. The results give:

$$\mu_K^* = \frac{\sum_i \sum_j E_{ij}\rho_{iK}\kappa_{jK}}{\left(\sum_i \rho_{iK}^2\right)\left(\sum_j \kappa_{jK}^2\right)}, \tag{9}$$

$$\alpha_{iK}^* = \frac{\sum_j (E_{ij} - \mu_K \rho_{iK}\kappa_{jK})\kappa_{jK}}{\rho_{iK}\left(\sum_j \kappa_{jK}^2\right)}, \tag{10}$$

$$\beta_{jK}^* = \frac{\sum_i (E_{ij} - \mu_K \rho_{iK}\kappa_{jK})\rho_{iK}}{\kappa_{jK}\left(\sum_i \rho_{iK}^2\right)}. \tag{11}$$

Given the values of $\rho_{iK}^{(s-1)}$ and $\kappa_{jK}^{(s-1)}$ from the $(s-1)$st iteration, we use Eqs. (9)–(11) to update $\theta_{ijK}^{(s)}$ at the sth iteration. Because updating α_{iK}^* only requires data for the ith juvenile, and updating β_{jK}^* only requires data for the jth variable, the resulting iterations are very fast. Given values for θ_{ijK}, the update formulas for ρ_{iK} and κ_{jK} are found by differentiating Eq. (7) wrt ρ_{iK} and κ_{jK}, setting the results equal to zero, and solving. This gives:

$$\rho_{iK}^* = \frac{\sum_j E_{ij}\theta_{ijK}\kappa_{jK}}{\sum_j \theta_{ijK}^2 \kappa_{jK}^2}, \tag{12}$$

$$\kappa_{jK}^* = \frac{\sum_i E_{ij}\theta_{ijK}\rho_{iK}}{\sum_i \theta_{ijK}^2 \rho_{iK}^2}. \tag{13}$$

The initial values of all the ρs and the κs are set in $(0, 1)$ (e.g., make them all equal to 0.5). Then, given values of $\theta_{ijK}^{(s)}$ and $\kappa_{jK}^{(s-1)}$, we use Eq. (12) to update $\rho_{iK}^{(s)}$, and similarly, given values of $\theta_{ijK}^{(s)}$ and $\rho_{iK}^{(s-1)}$, we use Eq. (13) to update $\kappa_{jK}^{(s)}$. Further details of the algorithm and suggestions for improving convergence can be found in ref. 32. At convergence, the estimated parameters for the kth layer are denoted by $\widehat{\mu}_k, \widehat{\alpha}_{ik}$, and $\widehat{\beta}_{jk}$, $k = 1, 2, \ldots, K$.

Software. The original plaid program can be downloaded from the website http://www-stat.stanford.edu/∼owen/clickwrap/plaid. An alternative version is available as BCPLAID in the *R* package biclust [50] based upon work by Turner *et al*. [46], who used instead a binary least-squares iterative procedure to estimate the plaid parameters.

4.5. The Plaid Model Applied to the Data on Juveniles

The plaid model was fitted to the data (6768 juveniles, 27 variables) on juvenile recidivism in an unsupervised learning mode, meaning that we did not include the response variable that specifically identifies juveniles who recidivate. The very few missing records were imputed by the mean for a continuous variable and by the modal category for a categorical variable. As part of the plaid setup, we required $\widehat{\mu}_k + \widehat{\alpha}_{ik}$ to have the same sign for all juveniles in the kth layer and $\widehat{\mu}_k + \widehat{\beta}_{jk}$ to have the same sign for all variables in the kth layer. The plaid model iterations terminated at 52 layers; see Table 2 for the plaid biclustering results. These layers were determined as descriptive tools to aid in identifying subsets of the data that could be useful in policy making. The number of juveniles in any layer ranged from 6 to 3357 and the number of variables in any layer ranged from 1 to 12. Of the 6768 juveniles, 110 were not members of any plaid layer.

4.6. Descriptions of Layers

For each derived layer, we computed the G_i statistic (Eq. (1)) for each juvenile, but this time the $\{X_j\}$ were defined using the results from the plaid analysis. To avoid any possible confusion for the reader, we emphasize that the G_i computed here uses a different definition of the $\{X_j\}$ than was used previously in the construction of G_i as one of the input variables to the plaid biclustering algorithm (see Table 1). Here, we would like to define X_j in such a way that it can be interpreted as a measure of the jth juvenile's strength of membership in the kth layer. The obvious choice would be to take it to be the effect $\widehat{\mu}_k + \widehat{\alpha}_{jk}$; however, with that definition, X_j can be either negative or positive, which violates the definition of G_i. Fortunately, because $\widehat{\mu}_k + \widehat{\alpha}_{jk}$ has the same sign for all j in the kth layer, the numerator and denominator of G_i will have the same sign. So, we take $X_j = |\widehat{\mu}_k + \widehat{\alpha}_{jk}|$. This definition of X_j produced a version of the G_i statistic that was used to see if there were spatial clusters of juveniles with strong or weak membership in the layer. This turned out to be very useful because, when visualizing thousands of points, it is difficult to detect a spatial pattern visually.

To illustrate some of the more interesting layers, the maps of layers 1, 3, 5, 6, 8, and 34 are given in Fig. 3. In each map, a juvenile's point is colored red (for a hotspot) if there exists a significant local cluster of high degree of membership in the layer, blue (for a coldspot) if there is a significant local cluster of low degree of membership in the layer, black if the juvenile is in the layer but not in a significantly high or low local cluster of membership, and light gray if the juvenile is not included in the layer. Recall that the neighborhoods are shown in Fig. 2.

Some of the layers had a strong clustering effect, which suggests neighborhood level causal mechanisms of recidivism, and these layers had lower-than-average recidivism rates; see Table 3 for the recidivism rates of all 52 layers. We see this, for example, in layers 1 (African-American neighborhoods that are generally working- and middle-class; recidivism rate 0.352), 6 (several disparate neighborhoods with one thing in common: a mix of

Table 2. Results from a plaid analysis of the juvenile recidivism data.

Layer	# Juveniles	# Variables	Variable names
1	3357	3	age, p_highsch, p_black
2	1494	12	jhismh, WhiteDum, Hispanic_Dum, Probation, sexoff, Prioroutofhomepl, ParenDeceased, ParentalCrime, victinj, PriorPersonalChgs, den_dr_sale, p_spanish
3	2279	2	age, DrugAbuse
4	1637	9	sexoff, ParenDeceased, WhiteDum, jhismh, Prioroutofhomepl, Probation, AlchoholAbuse, DrugAbuse, den_dr_sale
5	1989	2	LiveInstitution, InstantProperty
6	2385	2	victinj, InstantPerson
7	894	10	Hispanic_Dum, ParentalCrime, sibarr, sexoff, Prioroutofhomepl, InstantPerson, InstantProperty, p_spanish, den_dr_sale
8	1657	3	Juvdrgar, kcnt_1km, gi
9	443	11	ParSubAbuse, InstantProperty, InstantPerson, Hispanic_Dum, ParentalCrime, Juvdrgar, sexoff, WhiteDum, jhismh, p_spanish, den_dr_sale
10	213	11	InstantProperty, Juvdrgar, PriorPersonalChgs, ParenDeceased, Probation, WhiteDum, sexoff, Hispanic_Dum, Prioroutofhomepl, den_dr_sale, p_spanish
11	481	11	PriorPersonalChgs, Juvdrgar, victinj, ParentalCrime, InstantProperty, InstantPerson, Probation, WhiteDum, jhismh, sexoff, Prioroutofhomepl
12	391	3	p_highsch, LiveInstitution, sibarr
13	691	10	sibarr, InstantProperty, Hispanic_Dum, victinj, Juvdrgar, p_spanish, ParenDeceased, Probation, den_dr_sale, p-vacant
14	318	6	p_spanish, kcnt_1km, den_person, age, gi, p_highsch
15	641	2	AlcoholAbuse, ParSubAbuse
16	985	10	LiveInstitution, ParSubAbuse, InstantPerson, AlchoholAbuse, Juvdrgar, DrugAbuse, PriorPersonalChgs, Hispanic_Dum, ParenDeceased, sexoff
17	281	3	PriorPersonalChgs, DrugAbuse, LiveInstitution
18	1172	1	sibarr
19	836	10	ParSubAbuse, InstantProperty, sibarr, InstantPerson, ParentalCrime, victinj, ParenDeceased, sexoff, WhiteDum, Prioroutofhomepl
20	116	4	Age, p_highsch, WhiteDum, AlchoholAbuse
21	368	2	LiveInstitution, ParentalCrime
22	563	3	age, PriorPersonalChgs, p_highsch
23	335	4	kcnt_1km, gi, p_spanish, PriorPersonalChgs
24	81	12	Victinj, Hispanic_Dum, PriorPersonalChgs, ParentalCrime, Juvdrgar, sibarr, p_spanish, Probation, WhiteDum, sexoff, AlcoholAbuse, den_dr_sale
25	165	11	Hispanic_Dum, InstantProperty, PriorPersonalChgs, ParentalCrime, Probation, WhiteDum, p_spanish, ParenDeceased, Prioroutofhomepl, den_dr_sale, p_vacant
26	131	10	Hispanic_Dum, PriorPersonalChgs, victinj, Juvdrgar, LiveInstitution, Probation, Prioroutofhomepl, sexoff, sibarr, p_vacant
27	341	3	age, p_highsch, WhiteDum
28	263	10	Hispanic_Dum, InstantPerson, jhismh, victinj, PriorPersonalChgs, sexoff, WhiteDum, p_spanish, InstantProperty, p_vacant
29	536	4	p_black, gi, kcnt_1km, den_person
30	1079	8	LiveInstitution, victinj. AlcoholAbuse, Juvdrgar, sibarr, DrugAbuse, ParentalCrime
31	697	1	LiveInstitution
32	114	5	p_highsch, p_black, ParentalCrime, jhismh, ParSubAbuse
33	108	4	age, p_highsch, p_black, ParenDeceased
34	1202	2	DrugAbuse, AlcoholAbuse
35	285	1	ParSubAbuse
36	65	4	Hispanic_Dum, den_person, gi, kcnt_1km
37	39	6	p_spanish, Hispanic_Dum, den_person, gi, kcnt_1km
38	363	5	Age, Juvdrgar, den_person, p_black, p_highsch
39	114	5	p_spanish, gi, kcnt_1km, Hispanic_Dum, jhismh
40	208	11	InstantPerson, InstantProperty, Juvdrgar, Hispanic_Dum, sibarr, p_spanish, Prioroutofhomepl, sexoff, WhiteDum, ParenDeceased, AlcoholAbuse
41	226	8	Victinj, ParenDeceased, AlcoholAbuse, WhiteDum, InstantPerson, sexoff, Hispanic_Dum, Juvdrgar
42	315	2	Kcnt_1km, ParenDeceased
43	424	1	Probation

Table 2. Continued

Layer	# Juveniles	# Variables	Variable names
44	306	8	PriorPersonalChgs, Probation, Juvdrgar, ParenDeceased, ParSubAbuse, p_black, InstantProperty, sexoff
45	111	8	PriorPersonalChgs, Probation, WhiteDum, victinj, AlcoholAbuse, ParSubAbuse, sexoff, InstantPerson
46	192	4	p_spanish, Hispanic_Dum, gi, kcnt_1km
47	65	10	Prioroutofhomepl, ParentalCrime, sexoff, Probation, Hispanic_Dum, InstantProperty, ParenDeceased, InstantPerson, den_dr_sale, p_vacant
48	105	2	p_highsch, jhismh
49	347	3	Gi, ParentalCrime, jhismh
50	496	1	PriorPersonalChgs
51	24	6	age, den_person, sexoff, ParSubAbuse, White_Dum, p_highsch
52	6	5	Juvdrgar, p_highsch, age, den_person, Hispanic_Dum

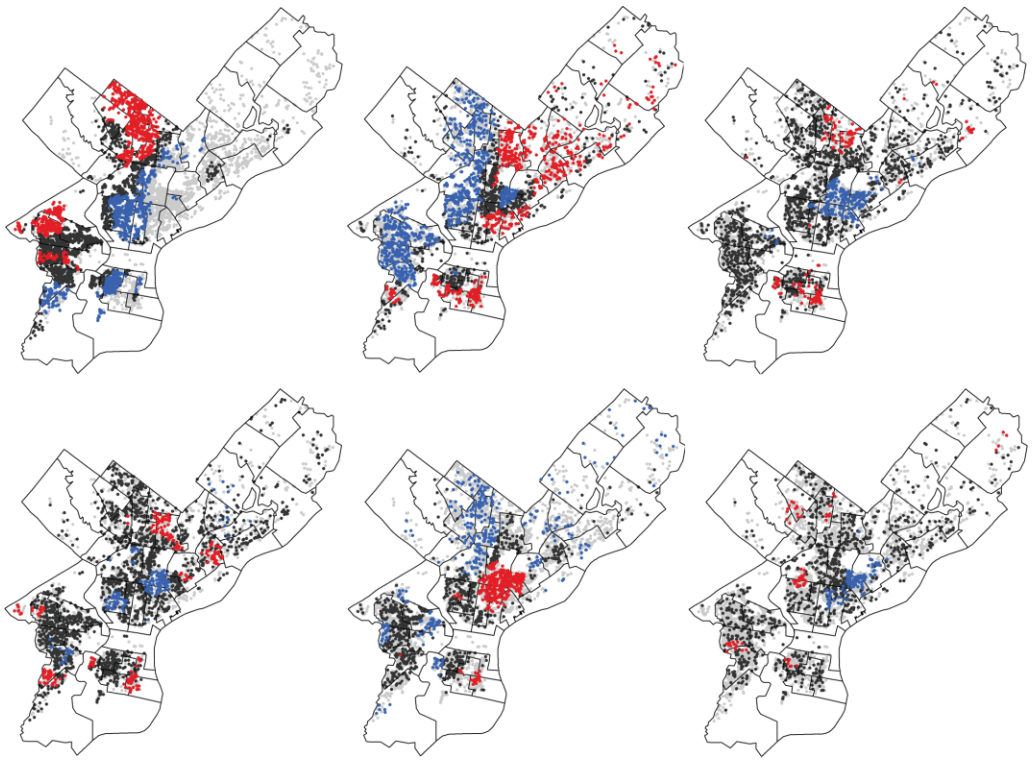

Fig. 3 Hotspot analysis of plaid layers 1 (top left), 3 (top center), 5 (top right), 6 (bottom left), 8 (bottom center), and 34 (bottom right). For each layer, red dots show hotspots, blue dots show coldspots, black dots show other juveniles in the layer, and gray dots show juveniles not in the layer. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

white and African-American residents; 0.334), and 29 (the juveniles reside almost exclusively in poor African-American neighborhoods; 0.330). Other layers exhibited little spatial clustering, which suggests individual- or family-level causal mechanisms of recidivism, and these layers had higher-than-average recidivism rates. We see this, for example, in layers 3 (all juveniles had a history of drug abuse; 0.427), 5 (large percentage of juveniles lived in an institution; 0.401), 8 (all juveniles had a prior drug arrest; 0.460), 15 (all juveniles had a parent with a history of substance abuse and a large percentage of the juveniles had a history of alcohol abuse; 0.410), 18 (all juveniles had a sibling with an arrest record; 0.388), 22 (almost all juveniles had prior personal charges; 0.455), 31 (all juveniles lived in an institution at the time of their instant offense; 0.428), 34 (all juveniles had a history of alcohol abuse and a high percentage had a history of drug abuse; 0.423), 43 (all juveniles were on probation at the time of their instant offense; 0.481), and 50 (all juveniles had prior personal charges and resided

Table 3. Recidivism rates for all 52 plaid layers. The overall recidivism rate is 0.387.

Layer	Recidivism ratio	Layer	Recidivism ratio	Layer	Recidivism ratio	Layer	Recidivism ratio
1	0.352	14	0.387	27	0.343	40	0.351
2	0.364	15	0.410	28	0.502	41	0.478
3	0.427	16	0.322	29	0.330	42	0.365
4	0.379	17	0.459	30	0.330	43	0.481
5	0.401	18	0.388	31	0.428	44	0.356
6	0.334	19	0.408	32	0.368	45	0.414
7	0.389	20	0.362	33	0.491	46	0.349
8	0.460	21	0.454	34	0.423	47	0.385
9	0.451	22	0.455	35	0.379	48	0.381
10	0.404	23	0.430	36	0.400	49	0.360
11	0.403	24	0.420	37	0.333	50	0.450
12	0.486	25	0.352	38	0.433	51	0.417
13	0.388	26	0.374	39	0.421	52	0.000

primarily in African-American neighborhoods; 0.450). For comparison purposes, we note that the recidivism rate over all 6768 juveniles was 0.387.

5. NONLINEAR MODELING OF PLAID LAYERS

So far, the plaid layers were constructed without regard to the recidivism status of each juvenile. The next step uses a nonlinear model to predict the recidivism rate for each plaid layer. In addition to the ProDES variables, we acquired data on community efficacy, socioeconomic character, and crime from a variety of sources, and we aggregated US Census Bureau tract data on rates for individual neighborhoods [25,31]. For this particular study, we focused on 436 legitimate predictors that might explain juvenile recidivism. Converting all the categorical variables to binary dummy variables expanded this number to 839 potential input variables. The 6768 juveniles were reduced to 6675 by removing all juveniles with more than 20% missing data. We also removed all variables with more than 30% missing data. All other missing data were imputed. From the remaining variables, we retained those that were identified for each plaid layer as the most significant predictors of recidivism. For each layer, we formed a 2×2 table for each variable separately (ignoring dependencies): the rows for a dummy variable were the two outcomes 0 or 1, while a continuous variable was split into high- and low-value categories; the columns reflected the state of recidivism (i.e., whether a juvenile recidivated within 6 months of program discharge); and the cells were the joint frequencies over all juveniles in that layer. A 'candidate' variable was one with a nonzero value of the usual chi-squared statistic χ^2. For each layer, we selected at most 40 of the candidate variables having the largest χ^2 values, dropping redundant variables that had been generated from other variables and which had very similar χ^2 values, and used the remaining $r \leq 40$ variables as input nodes to a neural network model [51].

Neural networks (see, e.g., ref. 32, Ch. 10) are parameterized multivariate statistical models for investigating nonlinear dependencies between the input and output variables that are too complicated for methods such as logistic regression or decision trees. The simplest 'feed-forward' type of neural network has a layer of r input nodes (X_m, $m = 1, 2, \ldots, r$), a single layer of t hidden nodes (Z_j, $j = 1, 2, \ldots, t$), and a layer of s output nodes (Y_k, $k = 1, 2, \ldots, s$). Let β_{mj} be the weight of the connection $X_m \to Z_j$ with bias β_{0j}, and let α_{jk} be the weight of the connection $Z_j \to Y_k$ with bias α_{0k}. Let $\mathbf{X} = (X_1, \ldots, X_r)^\tau$ and $\mathbf{Z} = (Z_1, \ldots, Z_t)^\tau$. Let $U_j = \beta_{0j} + \mathbf{X}^\tau \beta_j$ and $V_k = \alpha_{0k} + \mathbf{Z}^\tau \alpha_k$, where $\beta_j = (\beta_{1j}, \ldots, \beta_{rj})^\tau$ and $\alpha_k = (\alpha_{1k}, \ldots, \alpha_{tk})^\tau$. Then,

$$Z_j = f_j(U_j), \quad j = 1, 2, \ldots, t, \tag{14}$$

$$\mu_k(\mathbf{X}) = g_k(V_k), \quad k = 1, 2, \ldots, s, \tag{15}$$

where $f_j(\cdot)$, $j = 1, 2, \ldots, t$, and $g_k(\cdot)$, $k = 1, 2, \ldots, s$, are activation functions for the hidden and output layers of nodes, respectively. Putting these equations together, the value of the kth output node can be expressed as

$$Y_k = \mu_k(\mathbf{X}) + \epsilon_k, \tag{16}$$

where

$$\mu_k(\mathbf{X}) = g_k\left(\alpha_{0k} + \sum_{j=1}^{t} \alpha_{jk} f_j\left(\beta_{0j} + \sum_{m=1}^{r} \beta_{mj} X_m\right)\right), \quad k = 1, 2, \ldots, s, \tag{17}$$

and ϵ_k is the error term, which can be taken as Gaussian with mean zero and variance σ_k^2. The $\{f_j(\cdot)\}$ and $\{g_k(\cdot)\}$ are

taken to be nonlinear continuous functions with sigmoidal shape (e.g., logistic or tanh functions). The number t of hidden layers depends upon r, s, and the number of observations. A popular rule that allows easy repetitions of experiments is to assign $t = [(r + s)/2]$, the average of the number of input nodes and the number of output nodes, although there are other recommended guidelines.

The connection weights $\{\beta_{mj}\}$ and $\{\alpha_{jk}\}$ are initialized using randomly generated starting values and then estimated through an iterative gradient-descent optimization to minimize the error sum of squares,

$$ESS = \sum_{i=1}^{n} \sum_{k=1}^{s} (Y_{ik} - \mu_k(\mathbf{X}_i))^2, \tag{18}$$

on learning examples $\{(\mathbf{X}_i, \mathbf{Y}_i), i = 1, 2, \ldots, n\}$, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{is})^\tau$. If the error term ϵ_k is Gaussian, the resulting estimated weights are maximum-likelihood estimates. The backpropagation algorithm used for optimizing the connection weights is based upon an efficient computation of partial derivatives of an approximation function realized by the network. The learning data are fed through the network using possibly hundreds or thousands of iterations, calculating an output and adjusting the weights based upon their estimated influence on the observed errors on a learning example. The error correction function takes the partial derivative of the weight matrix to find minima for the outputs when compared to the output values and adjusts the weight and bias to calculate this derived value. This gradient-descent optimization is aimed at reducing the distance between the network's estimate and the actual output value. The effect is gradual and should improve with subsequent iterations until it converges to an optimal set of estimates for the outputs. Error functions have momentum and learning-rate parameters to control the adverse effects of erratic updates and local minima. The learning-rate parameter controls the amount of change that can occur in any given correction. The momentum parameter decreases the potential for drastic changes in the connection weights within the network; it takes into consideration the previous corrections by nudging the weights toward a single direction to avoid erratic updates. A weight-decay parameter is used to decrease the learning rate of the network slightly with each iteration; this stops the network from diverging from the output value and increases the network's performance.

Using our domain expertise and because trying to fit all 52 plaid layers by neural networks turns out to be highly computationally intensive, we restricted further attention to what were considered to be the most important layers. This gives the reader a view of a subset of the layers for illustrative purposes and enables us to pursue the modeling process without computational overload. Specifically, we carried out the modeling computations only for those layers in Table 2 that consisted of at least 400 juveniles and at most four variables. Layers with fewer juveniles were considered to be of little practical interest, while layers identified by large numbers of variables were considered to be too complicated for interpretation purposes. There were 13 layers that satisfied those restrictions; they were layers 1, 3, 5, 6, 8, 15, 18, 22, 29, 31, 34, 43, and 50.

In the previous section, we ran an unsupervised model in which plaid layers were obtained without using information on whether each juvenile recidivated during the duration of the study. Now, we fit a nonlinear supervised model, described by Eqs. (16) and (17), to the juveniles from each plaid layer with the recidivism status of each juvenile as the output variable, and then we use the resulting fitted model to predict the recidivism rate for that layer. Following the analyses described in the beginning of this section, the number of input variables for modeling each plaid layer was reduced to $r \leq 40$ input variables, where r was different for different layers. For each plaid layer, there were $s = 2$ output nodes (the juvenile recidivated or did not recidivate within six months of program discharge), and the number of hidden nodes was taken to be equal to $t = [(r + s)/2]$.

6. ACCURACY OF PREDICTIONS

One way of determining the accuracy of predictions of recidivism would be to fit the nonlinear model to the set of juveniles in each layer and then apply the fitted models to post-2004 juvenile delinquents; however, this was infeasible as such juvenile records were not in the ProDES database. Instead, we used the notions of learning set and test set to illustrate how well the models fit the data. First, we fit the neural network model to all the juveniles in each layer and then applied the fitted model for each layer to predict the recidivism status of those same juveniles. This yielded the apparent error rate or AER. We call this method the 'Learn' method because the split for each layer was 100% for the learning set and 0% for the test set. Second, we split up the juveniles in each plaid layer into two nonoverlapping groups by randomly assigning each one either to a learning set (80%) or to a test set (20%), we fitted the nonlinear model to the learning set from each layer, and then we applied that fitted model to determine the recidivism status of the juveniles in the test set only. This yielded the test-set error rate or TSER. We call this method the "Test" method. The neural network model for each layer was fit using WEKA's Multilayer Perceptron software package with learning-rate parameter set to 0.3, momentum parameter to 0.2, and the maximum number of iterations to 2000. We also computed the sensitivity and specificity rates for each layer (and over all juveniles, ignoring layer identification) for the ganypet,

xdrugs, xperson, and xproperty outcome variables. The definitions of sensitivity and specificity are

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}, \quad (19)$$

where TP is the number of true positives (i.e., # juveniles who were correctly predicted to recidivate), TN the number of true negatives (# juveniles who were correctly predicted not to recidivate), FP the number of false positives (# juveniles predicted to recidivate, but did not), and FN the number of false negatives (# juveniles predicted not to recidivate, but did). Thus, sensitivity measures the proportion of juveniles who are correctly predicted to recidivate, while specificity measures the proportion of juveniles who are correctly predicted not to recidivate. The results are listed in Table 4.

Because the modeling process was optimized for the learning set, we expect that the AER would be overly optimistic and, hence, smaller than the TSER for all layers. This held for all but one comparison, namely, layer 15 of xperson. If all types of recidivism are combined (ganypet), we see that only two of the 13 plaid layers have test-set error rates lower than that for the overall data set. When we specialize by type of recidivism, our results improve substantially. For drug-offense recidivism (xdrugs), eight plaid layers have test-set error rates lower than that for the overall data set; for personal-offense recidivism (xperson), six plaid layers have test-set error rates lower than that for the overall data set; and for property-offense recidivism (xproperty), eight plaid layers have test-set error rates lower than that for the overall data set. These results show that prediction of juvenile recidivism is substantially more

Table 4. Error rates, sensitivity rates, and specificity rates for (A) *ganypet* (any type of recidivism), (B) *xdrugs* (drug recidivism), (C) *xperson* (personal-offense recidivism), and (D) *xproperty* (property-crime recidivism) as derived from neural network modeling of each of the 13 selected plaid layers. For the 'Learn' entries, the test set was identical to the learning set. For the 'Test' entries, the juveniles were split randomly into a learning set (80%) and a test set (20%). AER estimates the apparent error rate for the 'Learn' juveniles and TSER estimates the test-set error rate for the 'Test' entries. The 'All' row represents the various rates over the entire data set.

	Error rates		Sensitivity rates		Specificity rates	
Layer	Learn (AER)	Test (TSER)	Learn	Test	Learn	Test
(A)						
1	0.199	0.365	0.559	0.282	0.933	0.811
3	0.158	0.456	0.675	0.523	0.966	0.561
5	0.201	0.429	0.639	0.556	0.906	0.581
6	0.141	0.357	0.613	0.329	0.982	0.802
8	0.153	0.491	0.733	0.542	0.945	0.480
15	0.020	0.508	0.973	0.321	0.984	0.613
18	0.068	0.509	0.834	0.458	0.994	0.510
22	0.228	0.420	0.874	0.455	0.689	0.702
29	0.136	0.414	0.690	0.471	0.949	0.643
31	0.026	0.425	0.960	0.561	0.985	0.585
34	0.051	0.454	0.917	0.471	0.972	0.605
43	0.195	0.429	0.708	0.634	0.895	0.512
50	0.173	0.500	0.863	0.412	0.798	0.596
All	**0.276**	**0.386**	**0.408**	**0.272**	**0.923**	**0.847**
(B)						
1	0.027	0.164	0.762	0.078	0.998	0.917
3	0.057	0.256	0.713	0.281	0.995	0.848
5	0.027	0.140	0.739	0.111	0.998	0.936
6	0.022	0.106	0.744	0.179	1.000	0.939
8	0.085	0.314	0.761	0.361	0.967	0.796
15	0.009	0.164	0.917	0.235	1.000	0.928
18	0.014	0.196	0.900	0.143	1.000	0.923
22	0.032	0.170	0.904	0.000	0.979	0.894
29	0.013	0.087	0.829	0.333	1.000	0.949
31	0.103	0.259	0.762	0.385	0.926	0.823
34	0.018	0.214	0.883	0.308	0.999	0.879
43	0.033	0.226	0.815	0.200	0.994	0.899
50	0.041	0.143	0.677	0.000	1.000	0.944
All	**0.077**	**0.213**	**0.520**	**0.138**	**0.985**	**0.899**

Table 4. Continued

Layer	Error rates		Sensitivity rates		Specificity rates	
	Learn (AER)	Test (TSER)	Learn	Test	Learn	Test
(C)						
1	0.049	0.137	0.542	0.057	0.997	0.958
3	0.060	0.122	0.535	0.159	0.980	0.956
5	0.047	0.175	0.591	0.088	0.994	0.950
6	0.090	0.136	0.338	0.100	0.985	0.976
8	0.067	0.082	0.231	0.000	0.990	0.993
15	0.092	0.086	0.333	0.083	0.977	1.000
18	0.088	0.113	0.270	0.095	0.991	0.967
22	0.102	0.125	0.164	0.067	0.998	1.000
29	0.082	0.125	0.196	0.000	0.996	0.958
31	0.042	0.166	0.588	0.000	0.998	0.928
34	0.029	0.147	0.705	0.143	0.999	0.922
43	0.109	0.167	0.193	0.000	1.000	1.000
50	0.120	0.204	0.132	0.000	1.000	1.000
All	**0.051**	**0.128**	**0.511**	**0.078**	**0.997**	**0.957**
(D)						
1	0.033	0.148	0.710	0.151	0.997	0.913
3	0.027	0.171	0.744	0.115	1.000	0.922
5	0.112	0.203	0.366	0.096	0.980	0.904
6	0.025	0.170	0.749	0.023	0.998	0.913
8	0.073	0.076	0.114	0.042	0.998	0.993
15	0.056	0.203	0.593	0.111	0.995	0.909
18	0.037	0.157	0.676	0.053	0.994	0.915
22	0.088	0.170	0.373	0.050	0.986	1.000
29	0.013	0.135	0.877	0.214	1.000	0.967
31	0.091	0.108	0.074	0.000	1.000	0.969
34	0.045	0.160	0.634	0.115	0.998	0.929
43	0.095	0.095	0.359	0.000	0.984	0.987
50	0.108	0.133	0.086	0.071	1.000	1.000
All	**0.067**	**0.160**	**0.416**	**0.058**	**0.996**	**0.942**

accurate by conditioning on type of recidivism rather than by combining all types into a single category.

Regarding the sensitivity and specificity measures, we see that the 'Learn' sensitivity rates are all higher than the 'Test' rates (as one would expect), but this is not always true for the specificity rates. In general, the specificity rates are very high, with rates for xdrugs, xperson, and xproperty being higher than those for ganypet. The sensitivity 'Test' rates, on the other hand, are very low, and for most layers the rates for xdrugs, xperson, and xproperty are lower than those for ganypet, which are not that high. It appears that predicting nonrecidivating juveniles is not difficult, but predicting juveniles who recidivate is extremely difficult, especially for drug recidivism, personal-offense recidivism, and property-crime recidivism.

7. DISCUSSION

In this paper, we presented the results of an investigation into the prediction of juvenile delinquency and recidivism in an urban setting. The initial stages of the statistical analysis of this large and complicated data set consisted of extensive data preprocessing and data-reduction work. In previous studies [25,31], we showed that because the relationships between many of the variables were quite weak, the use of hierarchical linear models, spatial econometric regression methods, and variable selection methods were not able to provide satisfactory explanations or predictions of juvenile recidivism. In this study, we suggest that the presence of nonstationarity in these data shows such one-step modeling to be very simplistic. To account for the nonstationarity in these data, we combined the use of a two-way, biclustering procedure, which subdivides the collection of juveniles into 'types', with nonlinear modeling applied to the different 'types' of juveniles. This two-step procedure enabled juvenile recidivism rates to be predicted quite well, although predicting which juveniles will recidivate is a much more complicated problem. Furthermore, by breaking down recidivism by offense categories, we were able to provide better prediction of juvenile recidivism rates, rather than by pooling all types of recidivism into a single category

and then trying to predict general recidivism. The research described in this paper is unusual in that few other published studies investigate juvenile recidivism by breaking down recidivating offenses by offense type, except for studies that investigate sex offenders who recidivate, and serious, chronic, and violent reoffenders. Although this study is focused on juvenile recidivism in Philadelphia, the methods employed here can be used for any urban setting.

Our measures of recidivism include a period of program participation and an additional 6 months. Although we did not include program effects in our models, we do know from our other research on the same data [25,31] that program attributes are unrelated to drug re-offending. However, for person and property offending, there appears to be a program effect that should be included in future work. Moreover, there is a large body of program evaluation research that supports the view that programs can reduce re-offending [52]. It would also be interesting to consider longitudinal analyses, such as changes in offending patterns with respect to age or experience, and any further analysis would benefit from including measures of parent–child relationships, as well as neighborhood-family interactions.

ACKNOWLEDGMENTS

The authors thank Anne Marie Ambrose, Commissioner of the Philadelphia Department of Human Services, and Philadelphia Police Commissioner Charles H. Ramsey for providing data for this project. The authors also thank Brian Lockwood for providing references to the juvenile recidivism literature, Art Owen and Heather Turner for helpful correspondence regarding plaid models, and three anonymous referees and the Editor-in-Chief each of whose editorial comments improved the paper substantially. This research was supported by Award No. 2006-IJ-CX-0022 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the Department of Justice.

REFERENCES

[1] R. J. Sampson, The embeddedness of child and adolescent development: a community-level perspective on urban crime, In Violence and Childhood in the Inner City, J. McCord, ed. Cambridge, Cambridge University Press, 1997.

[2] I. Ellen, T. Mijanovich and K. Dillman, Neighborhood effects of health, J Urban Affairs 23 (2001), 391–408.

[3] S. Messner and L. Anselin, Spatial analyses of homicide with areal data, In Spatially Integrated Social Science, M. F. Goodchild and D. Janelle, eds. New York, Oxford University Press, 2004.

[4] M. F. Goodchild and D. Janelle, Spatially Integrated Social Science, New York, Oxford University Press, 2004.

[5] S. A. Fotheringham, C. Brunsdon, and M. E. Charlton, Quantitative Geography: Perspectives on Spatial Data Analysis, London, Sage Publications, 2000.

[6] L. Anselin, Spatial Econometrics: Methods and Models, Dortrecht, Kluwer, 1988.

[7] K. Jones, Specifying and estimating multilevel models for geographical research, Trans Inst Brit Geograph 16 (1991), 148–159.

[8] S. A. Fotheringham, C. Brunsdon, and M. E. Charlton, Geographically Weighted Regression: The Analysis of Spatially Varying Relationships, Chichester, Wiley, 2002.

[9] B. Brown, Adolescents' relationships with peers, In Handbook of Adolescent Psychology, R. Lemer and L. Steinberg, eds. New York, Wiley, 2004, 363–394.

[10] J. A. Graber, J. Brooks-Gunn, and A. C. Petersen, Transitions Through Adolescence: Interpersonal Domains and Context, Hillsdale, NJ, Lawrence Erlbaum Associates, 1996.

[11] D. S. Elliott, W. J. Wilson, D. Huizinger, R. J. Sampson, and B. Rankin, The effects of neighborhood disadvantage on adolescent development, J Res Crime Delinquency 33 (1996), 389–426.

[12] R. J. Sampson, S. W. Raudenbush, and F. Earls, Neighborhoods and violent crime: a multilevel study of collective efficacy, Science 2777 (1997), 919–924.

[13] L. Steinberg, H. L. Chung, and M. Little, Reentry of young offenders from the justice system: a developmental perspective, Youth Violence Juvenile Justice 1 (2004), 1–18.

[14] W. J. Wilson, The Truly Disadvantaged: The Inner City, The Underclass, and Public Policy, Chicago, IL, University of Chicago Press, 1987.

[15] J. Brooks-Gunn, G. J. Duncan, P. K. Klebanov, and N. Sealand, Do neighborhoods influence child and adolescent development? Am J Sociol 99 (1993), 353–395.

[16] L. Kowaleski-Jones, Staying out of trouble: community resources and problem behavior among high-risk adolescents, J Marriage Family 62 (2000), 449–464.

[17] B. H. Rankin and J. M. Quane, Neighborhood poverty and the social isolation of inner-city African-American families, Social Forces 79 (2000), 139–164.

[18] B. H. Rankin and J. M. Quane, Social contexts and urban adolescent outcomes: the interrelated effects of neighberhoods, families, and peers on African-American youth, Social Prob 49 (2002), 79–100.

[19] R. L. Simons, C. Johnson, J. Beaman, R. D. Conger, and L. B. Whitbeck, Parents and peer group as mediators of the effect of community structure on adolescent problem behavior, Am J Community Psychol 24 (1996), 145–171.

[20] R. L. Simons, K. H. Lin, L. C. Gordon, G. H. Brody, V. Murry, and R. D. Conger, Community differences in the association between parenting practices and child conduct problems, J Marriage Family 64 (2002), 331–345.

[21] C. R. Shaw and H. D. McKay, Juvenile Delinquency and Urban Areas, Chicago, University of Chicago Press, 1942.

[22] L. E. Daigle, F. T. Cullen, and J. P. Wright, Gender differences in the predictors of juvenile delinquency, Youth Violence Juvenile Justice 5 (2007), 254–286.

[23] S. J. Funk, Risk assessment for juveniles on probation, Crim Justice Behav 26 (1999), 44–68.

[24] P. Mazerolle, Gender, general strain, and delinquency: an empirical examination, Justice Quart 15 (1998), 65–91.

[25] H. Grunwald, P. W. Harris, J. Mennis, Z. Obradovic, A. J. Izenman, and B. Lockwood, Predicting recidivism: analyzing the effects of individual, program, and neighborhoods with cross-classified hierarchical generalized linear models, Paper presented at the Annual Meeting of the American Society of Criminology, Atlanta, GA, 2007.
[26] A. Getis and J. K. Ord, The analysis of spatial association by use of distance statistics, Geograph Anal 34 (1992), 189–206.
[27] J. K. Ord and A. Getis, Local spatial autocorrelation statistics: distributional issues and an application, Geograph Anal 27 (1995), 286–306.
[28] T. Zhang, Limiting distribution of the G statistics, Stat Prob Lett 78 (2008), 1656–1661.
[29] J. E. Eck, S. Chainey, J. G. Cameron, M. Leitner, and W. E. Wilson, Mapping crime: understanding the hot spots, NIJ Special Report, National Institute of Justice, Office of Justice Programs, US Department of Justice, Washington, DC, 2005.
[30] N. Cressie, Statistics for Spatial Data, New York, John Wiley, 1991.
[31] J. Mennis, P. W. Harris, Z. Obradovic, A. J. Izenman, H. Grunwald, and B. Lockwood, The effects of neighborhood characteristics and spatial spillover on urban juvenile delinquency and recidivism, Prof Geograph 63(2) (2011), 1–18.
[32] A. J. Izenman, Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, New York, Springer, 2008.
[33] I. S. Dhillon, Co-clustering documents and words using bipartite spectral graph partitioning, In Proceedings of the 7th ACM SIGKCC International Conference in Knowledge Discovery and Data Mining, F. Provost and R. Srikant, eds. New York, ACM Press, 2001, 269–274.
[34] S. Busygin, G. Jacobsen, and E. Krämer, Double conjugated clustering applied to leukemia microarray data, Second SIAM ICDM, Workshop on Clustering High-Dimensional Data and its Applications, Arlington, VA, 2002.
[35] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, Spectral biclustering of microarray data: coclustering genes and conditions, Genome Res 13 (2003), 703–716.
[36] H. Cho, I. S. Dhillon, Y. Guan, and S. Sra, Minimum sum-squared residue co-clustering of gene expression data, Proceedings of the Fourth SIAM International Conference on Data Mining, 2004, 114–125. http://www.siam.org/proceedings/datamining/2004/dm04.php.
[37] Y. Cheng and G. M. Church, Biclustering of expression data, In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, vol 8, P. Bourne, M. Gribskov, R. Altman, N. Jensen, D. Hope, T. Lengauer, J. Mitchell, E. Scheeff, C. Smith, S. Strande, and H. Weissig, eds. San Diego, CA, AAAI Press, 2000, 93–103.
[38] A. Tanay, R. Sharan, and R. Shamir, Discovering statistically significant biclusters in gene expression data, Bioinformatics 18(Suppl. 1) (2002), S136–S144.
[39] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, Discovering local structure in gene expression data: the order-preserving submatrix problem, Proceedings of the Sixth Annual International Conference on Computational Biology (RECOMB-02), New York, ACM Press, 2002, 49–57.
[40] L. Lazzeroni and A. Owen, Plaid models for gene expression data, Stat Sinica 12 (2002), 61–86.
[41] J. Yang, H. Wang, W. Wang, and P. Yu, Enhanced biclustering on expression data, Third IEEE International Symposium on BioInformatics and BioEngineering, Loa Amamitos, CA, IEEE Computer Society, 2003, 321–327.
[42] G. Ambler and P. Green, Bayesian two-way clustering for gene expression data, EPSRC & RSS Workshop on the Statistical Analysis of Gene Expression Data, Wye College, University of Bristol, 2003. http://www.bgx.org.uk/wye/talks/Wye-Peter.pdf.
[43] Q. Sheng, Y. Moreau, and B. De Moor, Biclustering microarray data by Gibbs sampling, Bioinformatics 19 (Suppl. 2) (2003), S243–S252.
[44] A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel, Finding large average submatrices in high-dimensional data, Ann Appl Stat 4 (2009), 985–1012.
[45] H. L. Turner, Biclustering Microarray Data: Some Extensions of the Plaid Model, Ph.D. Dissertation, University of Exeter, UK, 2005.
[46] H. Turner, T. C. Bailey, and W. J. Krzanowski, Improved biclustering of microarray data demonstrated through systematic performance tests, Comput Stat Data Anal 48 (2005), 235–254.
[47] H. Turner, T. C. Bailey, W. J. Krzanowski, and C. A. Hemingway, Biclustering models for structured microarray data, IEEE/ACM Trans Comput Biol Bioinform 2 (2005), 316–329.
[48] J. Caldas and S. Kaski, Bayesian biclustering with the plaid model, IEEE Workshop on Machine Learning for Signal Processing (2008), 291–296. http://users.ics.tkk.fi/jcaldas/papers/plaid_mlsp08.pdf.
[49] A. Freitas, V. Afreixo, M. Pinheiro, J. L. Oliveira, G. Moura, and M. Santos, Improving the performance of the iterative signature algorithm for the identification of relevant patterns, Stat Anal Data Mining 4 (2011), 71–83.
[50] S. Kaiser and F. Leisch, A toolbox for bicluster analysis in R, Compstat, Proceedings in Computational Statistics, 2008.
[51] S. Wu and P. A. Flach, Feature selection with labelled and unlabelled data, In Proceedings of ECML/PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support, and Meta-Learning, M. Bohanec, M. Kasek, N. Lavrac, and D. Mladenic, eds. 2002, 156–167.
[52] P. Greenwood, Prevention and intervention programs for juvenile offenders, Future Child 18 (2008), 185–210.