

Daniel W. Goldberg
Dept. of Computer Science
University of Southern California
Los Angeles, CA 90089-0255
Email: daniel.goldberg@usc.edu

Xinbo Zhang
Dept. of Preventive Medicine
University of Southern California
Los Angeles CA 90089-9175
Email: xinbozha@usc.edu

Jennifer C. Marusek
Dept. of Preventive Medicine
University of Southern California
Los Angeles CA 90089-9175
Email: marusek_j@ccnt.usc.edu

John P. Wilson
Dept. of Geography
University of Southern California
Los Angeles, CA 90089-0255
Email: jpwilson@college.usc.edu

Beate Ritz
Dept. of Epidemiology
University of California, Los Angeles
Los Angeles CA 90095
Email: britz@ucla.edu

Myles G. Cockburn
Dept. of Preventive Medicine
University of Southern California
Los Angeles CA 90089-9175
Email: cockburn@usc.edu

DEVELOPMENT OF AN AUTOMATED PESTICIDE EXPOSURE ANALYST FOR CALIFORNIA'S CENTRAL VALLEY

Abstract: This paper details an automated methodology developed for associating historical pesticide exposure values to individuals in the agricultural region of California's Central Valley. In particular we will discuss the technical challenges that we had to overcome due to inaccuracies and inconsistencies in the underlying data sources used for both the estimation of the pesticides applied to an area as well as the reported land use of the area. Using a tiered approach based on classifying types of errors in the underlying data sources, we were able to create a framework that employs a "least possible error" assumption to minimize the error associated with an individual's calculated exposure values. Approaches such as that presented in this paper will be useful to other environmental modelers, epidemiologists, and spatial data users involved in similar studies that are incorporating any type of spatial data that may not be completely accurate, but still needs to be used with some level of confidence.

INTRODUCTION

Fundamental to many environmental epidemiology studies is the assignment of environmental exposure values to an individual and their subsequent use in spatial analysis. This information is typically obtained by intersecting the location of the individual and the geographic pattern of the environmental factor to be measured. The types, amounts, and time periods of these environmental factors being estimated and investigated by a study can vary greatly from hour-by-hour ultraviolet (UV) exposure (e.g., Rigel et al., 2003; Thieden et al. 2004), to daily exposure to smoke from a wild fire (e.g., Bowman and Johnston, 2005; Frankenberg et al., 2005; Johnston et al., 2006; Viswanathan et al., 2006), to yearly exposure to smog emissions (e.g., Bayer-Oglesby et al., 2005; Künzli et al., 2000; Nafstad et al. 2004), to lifetime exposure to particulate matter from living next to an airport (e.g., Steinmaus et al., 2004). Even though each of these studies would need to investigate multiple temporal periods, spatial areas, and environmental dosages, they all share one common aspect: the need to link regional exposure estimates to individuals. Calculating exposure estimates to the environmental factors in question require some form of reference data that maintain estimates of potential environmental measurements, from which individual estimates can be derived. Thus, critical to the spatial association (and any results derived there from) is the existence and usability of accurate environmental data describing the environmental factor to be investigated. However, having data at hand that is completely accurate and designed, *a priori*, to be useable is the exception, not the rule.

This situation stems from the many ways that these data sources are typically created. They can be the result of actual scientific measurements, reported data values, or calculated data values. For instance, it would be impossible to calculate potential individual UV exposure without either giving everyone a personal UV meter (which is expensive), or by having a data source that could tell you how much potential UV would have been in an area where that person was, at the time they were there. The latter, more realistic option, requires the existence of such a data source, but in fact stations which measure the full spectrum of the UV wavelength, both UVA and UVB are relatively rare. To provide insight on potential UVB dosages, recent research has shown that UVB estimates produced from UVA stations may be sufficiently accurate, and thus UVB estimates are being produced and utilized (e.g., Tatalovich et al., 2006).

However, this type of estimation raises a series of important questions that must be asked along and between all three axes: time, space, and dose. For example: Is the time period covered by the reference data relevant to the time period for which exposure is to be estimated? Is the dose described at an appropriate spatial resolution to realistically determine accurate exposure levels? Is the dose imputed from the data source relevant to personal exposure? Is the dose described at the correct temporal resolution to realistically determine accurate exposure levels?

Each of these issues have come to the forefront during a series of recent research investigations into the effect of pesticide exposure in California's Central Valley (e.g., Bell et al., 2001; Marusek et al., 2006; Nuckols et al., 2007; Reynolds et al., 2005; Rull et al., 2001; 2003; 2006), as well as elsewhere in the country (e.g., Brody et al., 2002; 2004; Ward et al., 2000). As these studies have progressed, they have attempted to improve upon calculated pesticide exposure estimates by improving upon the spatial and temporal specificity of the reference data used. In the first paper in this series, Bell et al. (2001) calculated exposure estimates utilizing the Pesticide Use Reports (PUR) (California Department of Pesticide Regulation, 2000), which link the type, amount in pounds, acreage applied, date, method, and locations for which regulated pesticides had been applied to the Public Land Survey System (PLSS) sections. Rull et al. (2001; 2003) investigated the exposure misclassification resulting from the spatial resolution of those one square mile sections (i.e., grid cells), and the likelihood that the pesticide may not disperse equally within the cells. They compared the approach of Bell et al. (2001) with one using the Land Use Reports (LU) (California Department of Water Resources, 2005) which describe the actual spatial geometry of the croplands in terms of what crops are grown, from which it can be determined what pesticides were applied. This data source allows one to increase the validity of the resulting pesticide exposure estimates because it relies on finer spatial and temporal resolution source data. Increasing the certainty and validity of the results even further, Nuckols et

al. (2007), overcame a limitation of Rull et al. (2001; 2003) by not aggregating seasonal crops into a single classification and assuming each was equally likely, thus improving the temporal certainty. Likewise, outside of California, other researchers have been examining the feasibility of calculating historical lifetime exposures by incorporating additional/ancillary datasets (e.g. Brody et al., 2002; 2004; Ward et al., 2000).

What we can see from this progression is that as new data sources become available (both current and historical), there is an opportunity to run exposure models over and over again to both determine previous misclassifications, as well as to create more accurate exposure assessments. Further, as new climatic models are developed that explain the dispersion of environmental factors such as pesticides, the ability to easily incorporate them and improve the validity further will be needed. Therefore, this paper presents an approach to automate these procedures.

THE TROUBLE WITH CANCER AND PESTICIDE EXPOSURE

There are many facts which complicate epidemiological investigations into the possible link between pesticide exposure and cancer. First, environmental exposures which result in cancer likely occurred a long time ago, or over a long period of time. This means that historical data are needed to determine if an environmental exposure is the cause of a particular type of cancer.

Second, when looking at ambient residential exposure to pesticide, no one knows what they were exposed to. It is impossible to ask study participants questions like “How much paraquat were you exposed to in the air between the years of 1980 and 2000”? Likewise, the cost of measuring these exposure levels in blood samples of large populations is prohibitive, and probably irrelevant to lifetime or historical exposures. Given these limitations, geographic information systems (GIS) are used increasingly to calculate the potential amount of pesticide that one may have been exposed to.

Third, both cancer and ambient exposure to pesticides are relatively rare. Therefore, large study samples are required to accurately assess whether or not a particular pesticide can be linked as the cause of a particular cancer. The resulting output from processing these data sets and producing an exposure level to multiple pesticides per subject, per year, will be multiple times larger than the input, which can cause problems with existing database software typically used to store these results. For example, Microsoft Access has a 2GB file limit which can be prohibitive in large population studies. With this in mind, consider the minimum data elements an exposure record would need to contain in order to enable the calculation of per-subject, per-year exposure estimates on a per-chemical basis are listed along with their data types and sizes (Microsoft Corporation, 2007) in Table 1, indicating that, at a minimum, 40 Bytes would be required. Noting that the PUR database contains data for at least the last 15 years, we now have 600 Bytes (40*15) of data for a single subject and chemical. Further, the PUR database contains over 850 unique chemicals, resulting in the possibility of 510 KB (600*850) records per subject. This would then limit a study population to 4,000 subjects before the 2GB limit is reached. Granted, not all subjects will be exposed to all chemicals, so most likely less than 850 chemicals will be associated with each subject, but it should be clear that the size of the resulting subject exposure output will be large, given the study requirements listed earlier (the need for historical exposure data for multiple chemicals and people).

TABLE 1
EXAMPLE SUBJECT EXPOSURE RECORD WITH DATA TYPES AND SIZES

| Field | Data Type | Size (Bytes) |
|-------------------|-----------|--------------|
| subjectID | INTEGER | 4 |
| locationDateTime | DATETIME | 8 |
| locationLatitude | FLOAT | 8 |
| locationLongitude | FLOAT | 8 |
| chemicalID | INTEGER | 4 |
| chemicalExposure | FLOAT | 8 |
| Total | | 40 |

Taken as a whole, we can see that when developing or running a pesticide exposure model we will be faced with a huge amount of data because of the sheer number of subjects needed and the amount of data needed per subject. Further, if we also consider that the geographic functions typically applied to these data are computationally expensive, the situation gets even more bleak. Therefore, an efficient process is needed to compute these individual exposure levels.

PROJECT DESCRIPTION

Given these challenges, the goal of this project was to develop an automated methodology that could reliably process large amounts of input data in the form of subject locations and pesticide application data, calculate historical exposure values, be easily extensible in terms of geographic regions, flexible in terms of spatial and temporal resolution of subject and pesticide application data, with the ability to control output function. To our knowledge, there exist no published reports detailing exactly how the implementation of a pesticide exposure model was created, and thus the work presented in this paper should prove insightful for others who wish to develop their own models for determining environmental exposure using similar data sources and techniques.

As a starting point for automating this functionality, we chose to model the two most accurate pesticide exposure models available at the time. When this project began, the models chosen were that of Bell et al. (2001) and Rull et al. (2001; 2003). As noted, the first model made use of the PUR data available from the state of California Department of Pesticide Regulation which links pesticide applications to PLSS one square mile sections, and the second made use of the higher resolution LU data from the California Department of Water Resources.

PUR DATABASE DEFINITION

The PUR database can be considered as a set of n individual records, app_i , each representing a single pesticide application. The total database, or combined set of all applications, $[Apps]$, is simply the union of all individual applications. The size of this set, $|[Apps]|$, is equal to the number of application records, n .

$$[Apps] = \bigcup_{i=1}^n app_i \quad (1)$$

$$|[Apps]| = \bigcup_{i=1}^n app_i \quad (2)$$

Each of the individual applications app_i , within $[Apps]$ can be represented as a vector describing the attributes of interest; t , the date and time of application; c , the chemical applied; p , the pounds applied; a , the acreage applied to; and l , the location (PLSS ID) where it was applied.

$$app_i = \langle t, c, p, a, l \rangle \quad (3)$$

From each individual application record, app_i , the density of a particular chemical, c , applied during a particular application, $density(l, app_i(c))$, within a PLSS grid cell, l , can be computed by dividing the pounds applied, $app_i(p)$, by the application acreage, $app_i(a)$. In the case of the PUR database, $[Apps]$, each application, app_i , corresponds to the application of a single chemical within a single grid cell, l , so $density(l, app_i(c))$, can be rewritten simply as $density(app_i)$. This should be distinguished from the case where a single application app_i , represents the application of a set of multiple chemicals. In this latter case, the equations developed and used throughout this paper would need to be altered.

$$density(app_i(c)) = \frac{app_i(p)}{app_i(a)} \quad (4)$$

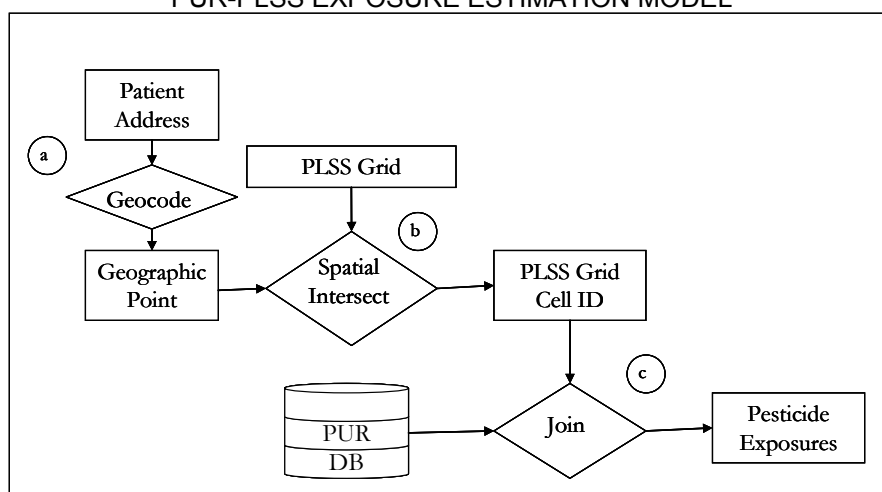
$$density(app_i) = density(app_i(c)) \quad (5)$$

It should be noted at this point that this density $density(app_i)$ of chemical application is associated with the entire area within the one square mile PLSS grid cell, l . From the PUR data alone, it is impossible to determine where within the cell the application occurred.

THE PUR-PLSS EXPOSURE MODEL

A generalized workflow diagram representing the processing steps necessary to complete the Bell et al. model (2001) in Figure 1. In step a) an address is turned into a spatial location by the process of geocoding. In b), this spatial point is then intersected with a spatial data layer containing geographic polygons representing the one square mile PLSS grids. In c), the PLSS grid cell ID of the intersected PLSS grid cell and the density of chemical application as defined in the PUR database are linked using a relational database join with the PLSS grid cell ID of the PLSS sections from b) as the key.

FIGURE 1
PUR-PLSS EXPOSURE ESTIMATION MODEL



THE PLSS GRID CELL SECTION

Because the PUR data has a maximum spatial resolution of one square mile, Bell et al. (2001) were forced to restrict their analyses of exposure to “narrow” and “broad” classifications, as shown in Figure 2. The black dot represents the location of a subject’s geocoded address, and narrow means the exposure levels were used from the single PLSS grid cell that a person lived in (cell with horizontal lines). Broad means this particular cell and the eight surrounding cells (with vertical lines) were used.

To formulate a notation for this phenomenon, we can first observe that the PLSS grid is essentially a raster data layer, or set $[G]$, composed of p columns and q rows, with each individual grid, $G_{x,y}$, being defined in terms of its relative position along the x and y axes. This is depicted graphically in Figure 3, where we have arbitrarily made the origin of the coordinate system the bottom left corner, and mathematically in Equation 6.

$$[G] = \bigcup_{x=0}^p \bigcup_{y=0}^q G_{x,y} \quad (6)$$

FIGURE 2
 BELL ET AL. (2001) EXPOSURE CLASSIFICATION; NARROW - HORIZONTAL, AND BROAD - VERTICLE

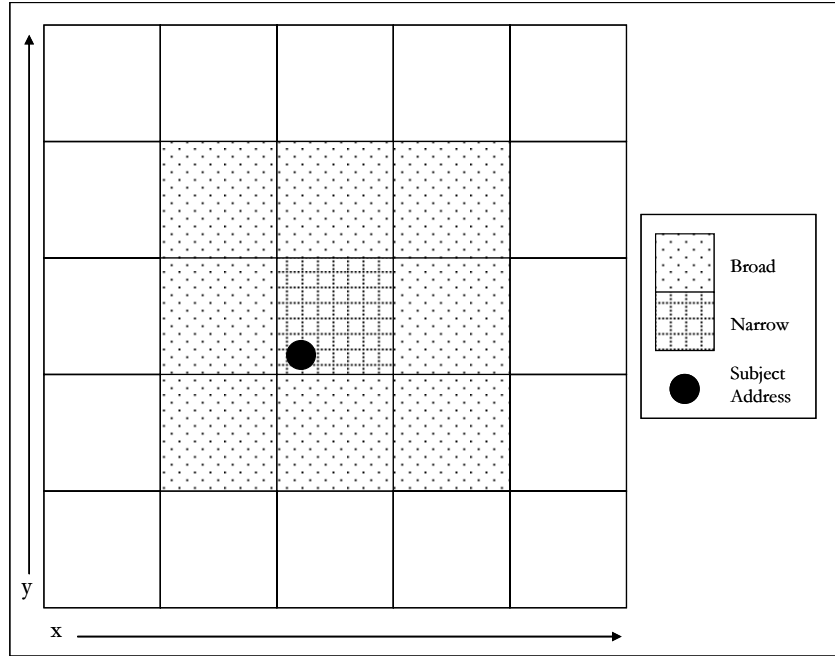
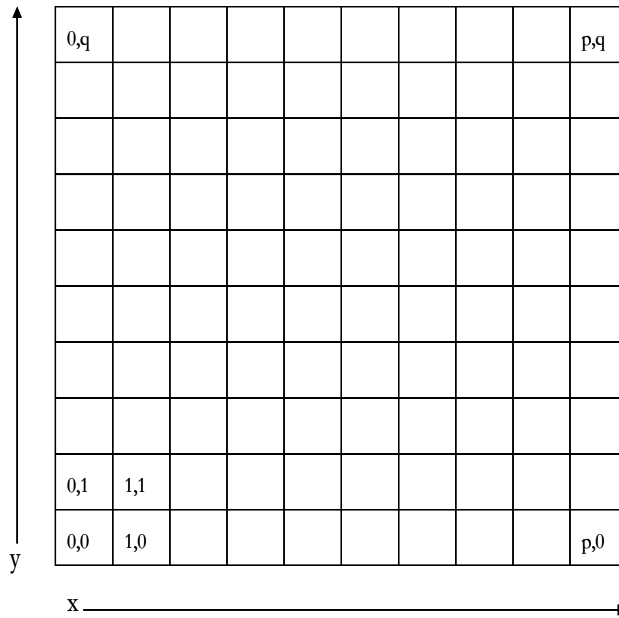


FIGURE 3
 PLSS GRID RASTER GENERALIZATION



With this notation in hand, we can define a subset of $[G]$, $[G_s]$, representing the set of PLSS grid cells to be included in the PLSS-PUR exposure estimation. We can further define an iterative

function (Equation 7) to uniformly create $[G_s]$, for both the narrow and broad cases. This is accomplished by incorporating a scale factor, s , into the grid cell selection equation, which also takes as input the total PLSS grid, $[G]$, and the grid cell where the residence is, $G_{x,y}$, determined through the spatial intersection of the geocoded address point and the PLSS grid layer. A pseudo-code implementation (written in Java/C# syntax) is shown in Listing 1.

$$[G_s] = \text{GetCells}([G], G_{x,y}, s) \quad (7)$$

LISTING 1
IMPLEMENTATION OF SCALABLE PLSS GRID CELL SELECTION ALGORITHM

```

cellList GetCells(grid g, cell c, int scale){
    cellList ret = null;
    if(c!=null){
        ret = new cellList();
        if(scale > 0){
            int min = scale * -1;
            int max = scale;
            for (int i=min; i<=max; i++){
                for (int j=min; j<=max; j++){
                    if (i>=0 && j>=0 && i<g.rows && j<g.cols){
                        cell temp = g[c.x + i, c.y + j];
                        ret.add(temp);
                    }
                }
            }
        }
        else{
            ret.add(c);
        }
    }
    return ret;
}

```

For Bell et al. (2001), $s = 1$, i.e., include 1 additional cell in every direction, for the broad case and $s = 0$, i.e., include 0 additional cells in every direction for the narrow case. By defining an algorithm to select the appropriate PLSS grid cells at any scale, we allow for different definitions of broad and narrow to be tested at different resolutions.

THE PUR-PLSS EXPOSURE ESTIMATION

The list of PLSS grid cells, $[G_s]$, generated from Equation 7, can be used to calculate $exp(c)$, the PUR-PLSS based exposure estimate, $exp()$, for a particular chemical, c . This calculation was based simply on the overall density of the application of the particular chemical, c , within each of the individual grid cells in $[G_s]$, as listed in the PUR database, $density(app(c))$ in Bell et al. (2001) because of the limitation of the PLSS resolution.

To accomplish this, we will define a relational database query, $getCellDensity(G_i, c)$, which takes a PLSS grid cell (PLSS ID) and chemical, and performs two tasks. The first is to select the correct individual pesticide application, app , from the total set of pesticide applications [Apps], based on its attributes $\langle t, c, p, a, l \rangle$, in particular, the identifier of the PLSS grid cell (PLSS ID) in which the application took place, l , and the chemical used, c . The second task is to calculate and

return the density of the chemical used in the application, $density(app(c))$. An SQL implementation for this function is illustrated in Listing 2.

LISTING 1
IMPLEMENTATION OF SCALABLE PLSS GRID CELL SELECTION ALGORITHM

```
SELECT (Pounds/Acres)
FROM PUR
WHERE (
    Chem=ChemID and
    PLSS=PLSSID
)
```

Thus, $exp(c)$, is calculated as the summation of densities, $density(app(c))$, of the applications of chemical c , which took place in the each of the m PLSS grid cells, in $[G_s]$.

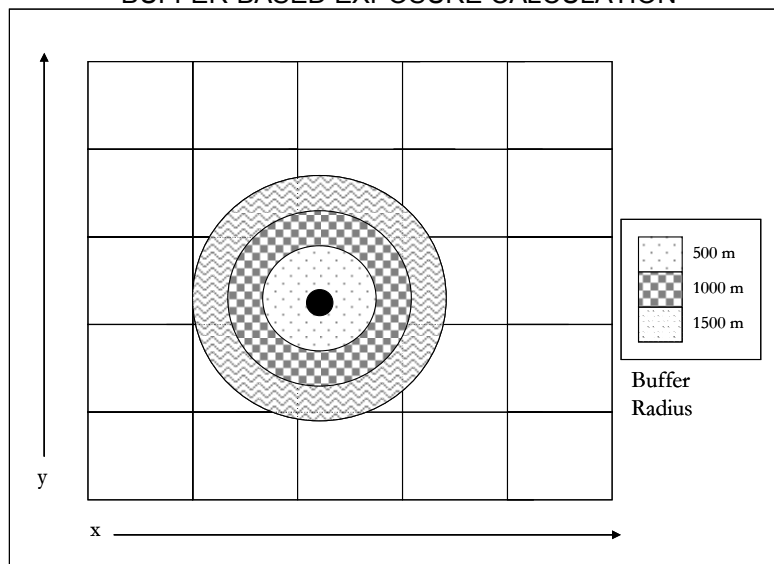
$$exp(c) = \sum_{i=0}^m getCellDensity(G_i, c) \quad (8)$$

It should be obvious that the exposure calculated with Equation 8, will overestimate the potential exposure values, and is not a true representation of a person's actual exposure. Pesticides are rarely dispersed evenly across an entire one square mile PLSS grid cell, and the density information from the PUR database cannot describe where in the cell it was applied. This is because, even though the PUR database reports the acreage that a chemical was spread over, from the PLSS data it is impossible to tell where within the one square mile it was applied, with every place being as likely as every other place.

EXPOSURE CALCULATION WITH BUFFERED GEOGRAPHIC PLSS AREAS

Instead of simply using a broad/narrow exposure classification scheme which assumes that an individual is exposed to the full density reported for each PLSS grid cell they are in (narrow exposure), or that cell plus the eight which surround them (broad exposure), it has become common to use one or more catchment areas (spatial buffers) around a point to derive a better exposure estimate, as depicted in Figure 4. This topic was not explored in Bell et al. (2001), and we offer it here as the next logical extension to their work.

FIGURE 4
BUFFER-BASED EXPOSURE CALCULATION



In this case, after applying a geospatial “buffer” to create a circular area of a particular radius around the geocoded point as in Figure 5a), a geospatial “clip” can be used to cut the PLSS grid cells into just the sections which fall into the buffered region as depicted in Figure 5b). The result of the spatial clip is a set of m partial grid cells (spatial polygons) $[P]$ from which a more precise chemical exposure density can be calculated. An updated workflow taking these new operations into account is depicted in Figure 6.

FIGURE 5
SPATIAL BUFFER AND CLIP OF PLSS GRID CELLS

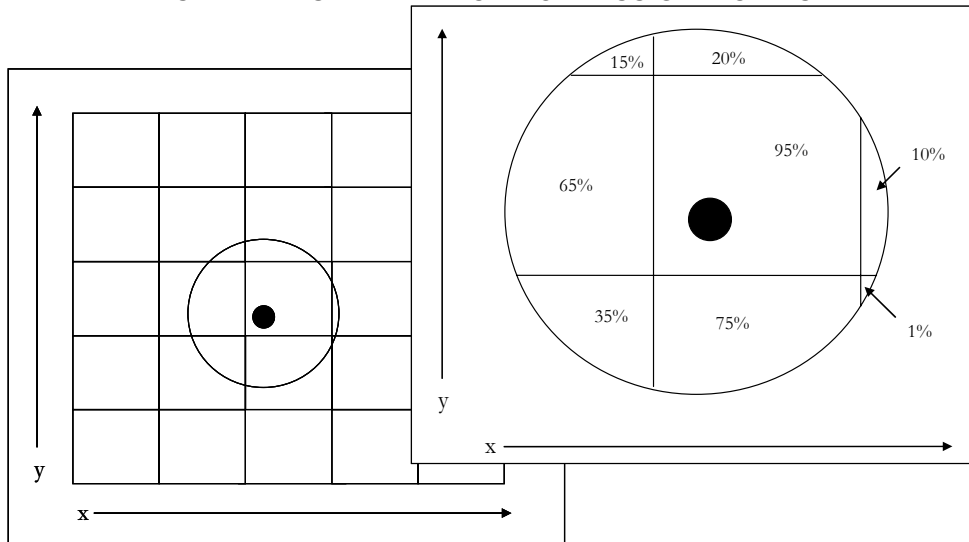
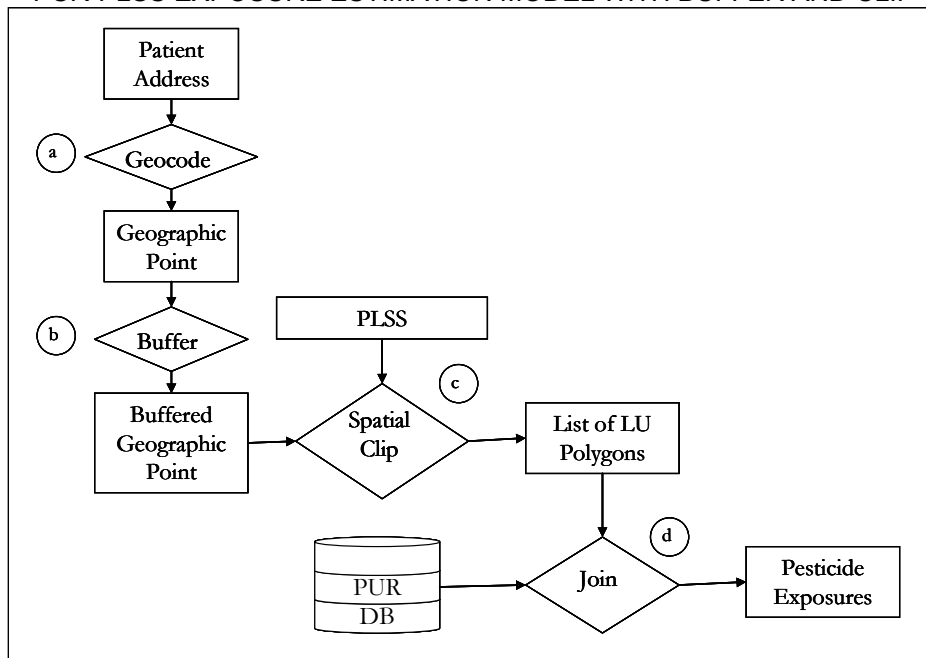


FIGURE 6
PUR-PLSS EXPOSURE ESTIMATION MODEL WITH BUFFER AND CLIP



By dividing the original density returned from the PUR database from the $getCellDensity(G_i, c)$ function, by the proportion of the area of the cell, P/mi^2 , the exposure estimate may more accurately represent the chemical application that a person might have been exposed to. This new exposure calculation is portrayed in Equation 9.

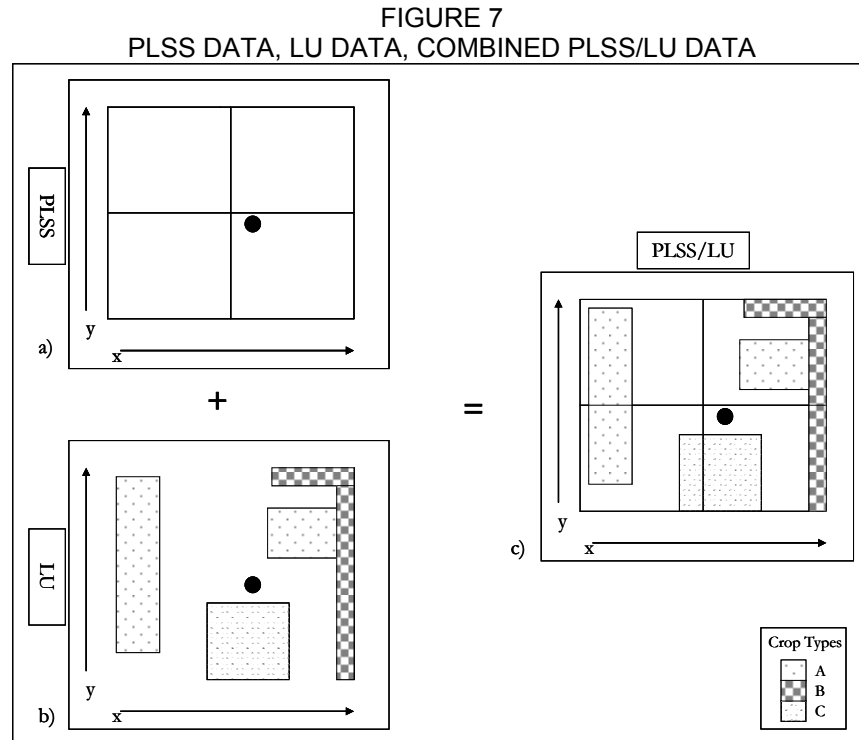
$$\text{exp}(c) = \sum_{i=0}^m \frac{\text{getCellDensity}(G_i, c)}{P / \text{mi}^2} \quad (9)$$

It should be noted that the exposure estimates produced with this new method still assume that the pesticides was applied at a constant density throughout the entire one square mile PLSS grid cell. This new method simply tries to obtain a more accurate proportion of the pesticide that a person was exposed to by taking an exposure equal to the proportion of the area of the PLSS grid cell they were potentially exposed to. At this point, no other information exists regarding the location of the application within the PLSS grid cell.

THE PUR-LAND USE EXPOSURE MODEL

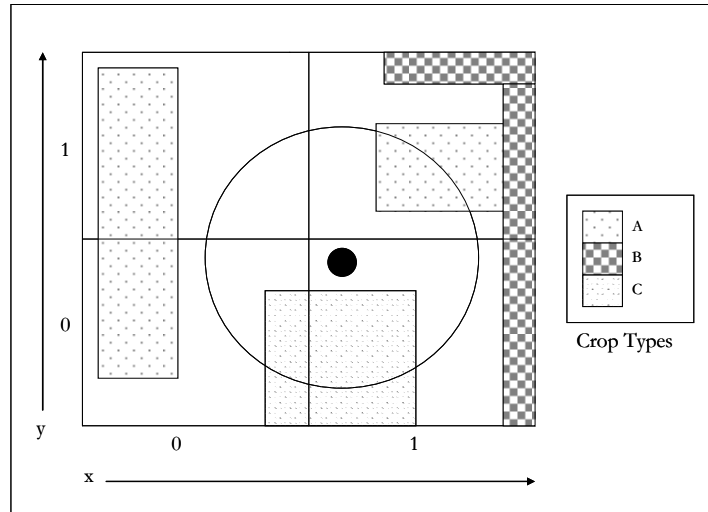
Rull et al. (2001; 2003) realized the assumption that the pesticide application occurs at a constant density across the PLSS grid cell is false (pesticides are not typically applied at a constant density across an entire one square mile PLSS grid cell), and inconsistent with the data in the PUR database which states the actual acreage the chemicals were applied to.

In contrast to the low resolution of the PLSS grid cell (Figure 7a)) the LU maps depict where within the one square mile PLSS grid cells the actual crop fields are located (Figure 7b)). These LU maps contain polygons with attributes describing which crops are grown in that field from surveys conducted by the State of California on a county-by-county basis, once every 7-10 years. However the polygons lack the PLSS grid cell IDs needed to link them directly to the PUR database. These two data sets can be spatially “joined” to associate these PLSS grid cell IDs with the crop field polygons (Figure 7c)), as accomplished by Rull et al (2001; 2003).



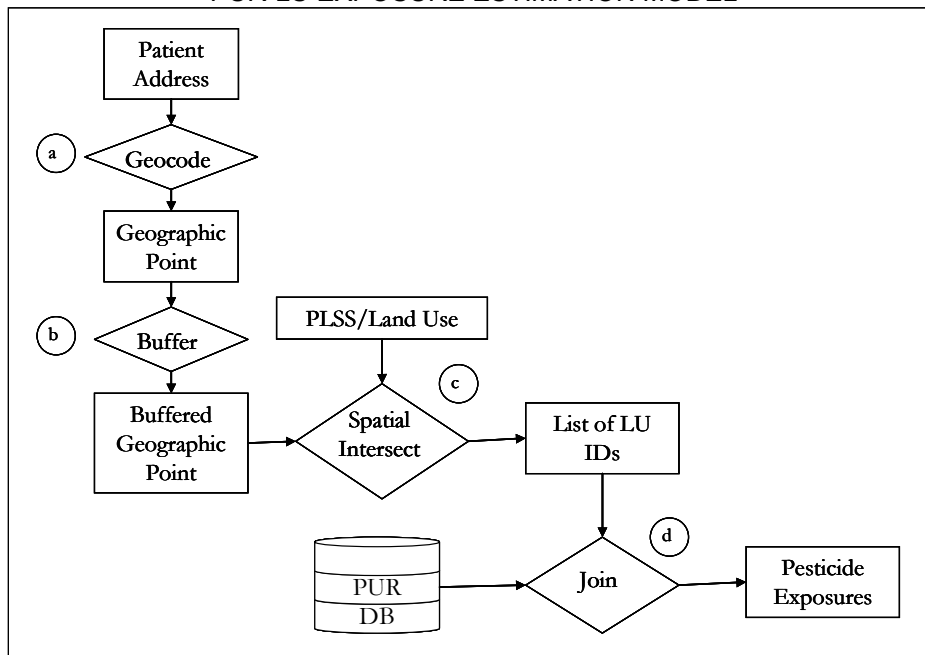
Rull et al. (2001; 2003) used these more accurate spatial distributions in conjunction with the buffered geocoded addresses to determine individual pesticide exposure as depicted in Figure 8. In this hypothetical example, this individual would be classified as exposed to chemical a in grid cells [1,1] and chemical c in grid cells [0,0] and [0,1].

FIGURE 8
 EXAMPLE EXPOSURE CLASSIFICATION BASED ON PLSS/LU MAP WITH BUFFERED
 GEOCODED ADDRESS



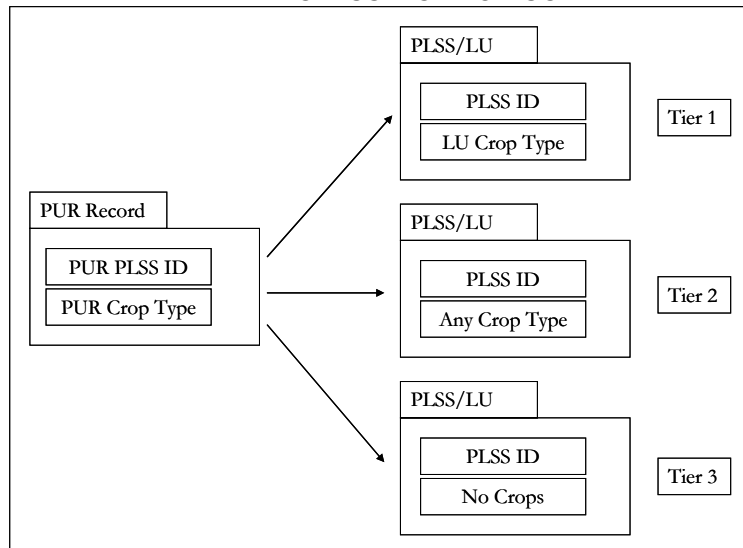
In Figure 9, the workflow from Figure 6 has been updated to account for the observation that the spatial locations of croplands from the LU data can be used to more accurately predict whether or not a subject should be classified as exposed to a chemical. The steps are the same as in Figure 6, except that the LU data are spatially intersected at step c) and used to more accurately calculate potential exposure values in step d).

FIGURE 9
 PUR-LU EXPOSURE ESTIMATION MODEL



Rull et al. (2001; 2003) developed a hierarchical scheme to account for incorrect data within and inconsistencies between the three data sets they used: the PLSS grid cells, PUR database, and LU polygons, as shown in Figure 10. These tiers represent level of uncertainty between the PUR data listing which chemicals were dispersed where, in what quantities, and on what crops, versus what the surveyed LU assignments were and how their temporal accuracy degrades over time as land, and in particular crop fields, are repurposed over time.

FIGURE 10
THREE TIER CLASSIFICATION SCHEME



A tier 1 match occurs in the case when the PUR record can be matched to a PLSS grid cell with the correct PLSS ID, and the PUR crop type exists within the PLSS/LU intersected layer. A tier 2 match occurs when the PLSS grid cell can be matched, but the PUR crop type does not exist in the PLSS/LU intersected layer. A tier 3 match occurs when the PLSS grid cell can be matched, but there are no crops within the PLSS/LU intersected layer.

For a tier 1 match, the chemical density from the PUR database is assumed to have been applied to only land from the PLSS/LU layer composing the fields of the particular crop types listed in the PUR database. For a tier 2 match, the chemical PUR density is assumed to have been applied to all of the crop fields within the PLSS/LU layer, taking into account that crop fields may have been repurposed in the years between LU surveys, and any of the crop fields are equally likely targets for pesticide application. For a tier 3 match the density is assumed to have been applied to the entire one square mile PLSS grid cell, because there is no other information about where within the cell it could have been applied.

It should be noted that even at the highest resolution match, the spatial resolution of the PUR location (PLSS ID) and the locations of the typed crop fields (LU) are not enough to determine sub-field application areas. For instance, a pest may strike a small section of a field to which a chemical is then applied, but it is impossible to determine this from the available data. Thus, for tiers 1 and 2, if multiple fields of the same crop (tier 1) or different crops (tier 2) are within a PLSS grid cell, it is not known which one it was applied to, so the most conservative assumption treats them all as equally likely. Further, the actual location within a field is not known, so again the most conservative assumption is to treat the entire area (of possibly more than one field) as an equally likely target for application. Therefore, the density calculations reflect this uncertainty by dispersing the density of the chemical from the PUR database (pounds/acres) over the whole area defined as the match area, at each of the tiers; total area of LU crop field of particular type within PLSS cell for tier 1, total area of all LU crop fields of any type within PLSS cell for tier 2, and total area of PLSS grid cell (one square mile) for tier 3.

An example LU polygon layer joined with the corresponding PLSS grid is depicted in Figure 11, and illustrates the various options (i.e., tier). Several examples from the PUR database for a few of these grid cells are shown in Table 2 which lists the chemical, the grid cell it was applied to, and the pounds and acres of the application. Table 2 also shows the tier the chemical was matched to along with the total number of fields of the particular crop type within the PLSS grid

cell and the number of polygons produced of the particular crop type within the cell after the buffer around the subject has been intersected with the LU polygons.

FIGURE 11
EXAMPLE PLSS/LU DISTRIBUTION

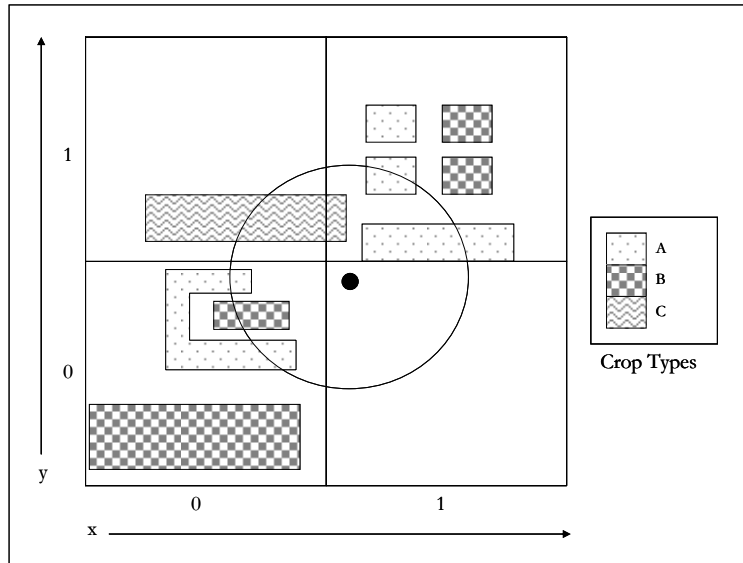


TABLE 2
EXAMPLE PUR-LU CHEMICAL APPLICATION TIERS

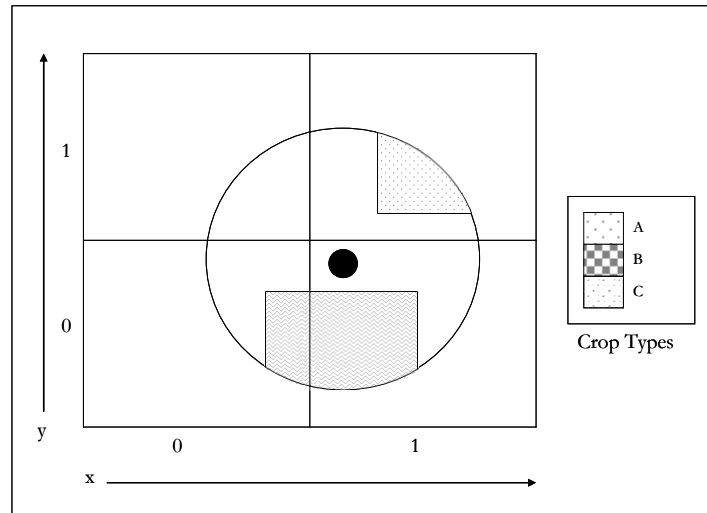
| Row | Chemical | Cell | Crop | Pounds | Acres | Tier | Fields | Intersects |
|-----|----------|-------|------|--------|-------|------|--------|------------|
| 1 | a | [1,0] | A | 25 | 300 | 3 | 0 | 0 |
| 2 | c | [0,0] | D | 15 | 30 | 2 | 3 | 3 |
| 3 | b | [0,0] | B | 10 | 100 | 1 | 2 | 1 |
| 4 | a | [0,0] | A | 20 | 60 | 1 | 1 | 2 |

For row 1, the PUR database reports that chemical a was distributed on crop type A in PLSS cell [1,0], but the LU lists no crops within this cell, resulting in a tier 3 match. For row 2, the PUR database reports that chemical c was distributed on crop type D within PLSS cell [0,0], but the LU does not report the crop type D as being within that cell, resulting in a tier 2 match. Additionally, there are three fields in PLSS cell [0,0], and the intersection with the individual's buffer results in three intersected polygons being produced as independent spatial geometric objects. In row 3, the PUR database reports that chemical b was applied to crop type B, which exists within the LU data for the correct PLSS cell resulting in a tier 1 match, with a total of two crop type B fields, and one intersect being created. Row 4 is also a tier 1 match, with one crop type A field in the PLSS cell and 1 intersect being created

EXPOSURE CALCULATION WITH LU CROP POLYGONS

The buffer-based approach we outlined as an extension to the work of Bell et al. (2001), was recently applied as an extension to the work of Rull et al. (2001; 2003) by Nuckols et al. (2007). The densities used to calculate exposure estimates were calculated as proportions of the total LU polygon sections created by the intersection of the LU data layer and the buffered geocoded address as in Figure 12, derived from the LU depicted in Figure 8.

FIGURE 12
 BUFFER-BASED PUR/LU EXPOSURE ESTIMATION

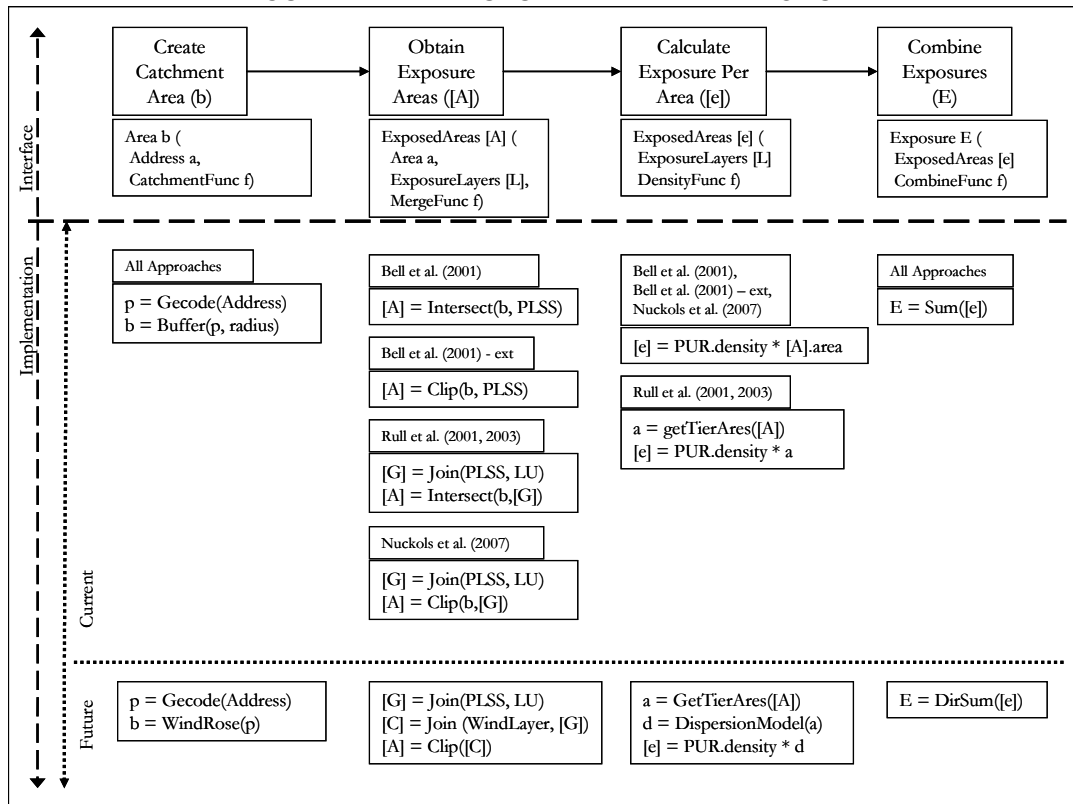


GENERALIZATION OF THE EXPOSURE ESTIMATION PROCESS

What should be clear at this point is that there are a fundamental series of data sources and operations required for any exposure assessment study. A breakdown of the four fundamental components of the exposure estimation process is modeled in Figure 13. These four operations are 1) create a catchment area, 2) obtain the exposure areas, 3) calculate the exposure per area, and 4) combine each individual exposure into a total exposure. This workflow is of course, extremely generalized, and has been specifically designed to encompass the four exposure estimation models described thus far which are (in order or increasing complexity), Bell et al. (2001), Bell et al. (2001) extended, Rull et al (2001; 2003), and Nuckols et al. (2007).

Figure 13 describes each of the four steps in terms of their inputs and outputs, with each step taking both data and functions (delegates) as parameters. Modeling the process in this fashion (using function delegates) is similar to using other software engineering techniques such as interfaces and should be implementable in most modern programming languages. An implementation in this manner affords the flexibility to define, research, and implement new methods for calculating exposure assessments as more accurate data sources become available or new environmental models are developed describing the atmospheric processes on pesticide application (e.g., wind rose data).

FIGURE 13
GENERALIZATION OF THE EXPOSURE MODELLING WORKFLOW AS INTERFACES WITH
CURRENT AND FUTURE IMPLEMENTATIONS



THE SCALABILITY PROBLEM

As we stated earlier, a major difficulty in performing epidemiological studies where cancer is the outcome and ambient pesticide exposure is the environmental factor is that the amount of data which needs to be processed is massive. The size of the subject population needs to be extremely large because cancer is rare, and the number of exposure estimates per subject will also be extremely large because there are a large number of chemicals to derive exposures for, and when deriving historical exposures, values will be needed for as many years as possible.

In the three studies available performing similar exposure estimations, the sizes of the populations used were relatively small; Bell et al. (2001): 684, Rull et al. (2001; 2003): 200, Nuckols et al. (2007): 577; all being 'convenience samples'. The reason for this is as Rull et al. state, "this approach was too computationally intensive for a simulation exercise involving more than 77,000 parcels" (Rull et al. 2003, pp 1588). In Figures 14 and 15, we can see the first part of the problem. Note that only a discussion of the methods of Rull et al. will be described, as implementation details for Bell et al. (2001) and Nuckols et al. (2007) have never been published, but a similar methodology is assumed.

Rull et al. (2001, 2003) performed portions of the processing as spatial processes within a GIS and portions as database processing within a relational database (Microsoft Access), as in Figure 14a. The spatial processes they performed within a GIS (ESRI ArcView), as shown in Figure 15 are, geocoding (Figure 15a), buffering (Figure 15b), and spatial intersection (Figure 15c). From here, the relational database tables describing the intersected spatial geometries were imported into Microsoft Access and the remainder of the processing was performed as "an elaborate series of queries" (Ritz, 2007, personal communication). The downside of this method is that if any changes are made to the process, the entire process needs to be run again from start to finish, e.g., changing the buffer shape or size. Also, the scalability of the process is equal to the

scalability of Microsoft Access, and the spatial indexes and capabilities of a GIS are not used for spatial operations, e.g. selection or join.

FIGURE 14
TYPES OF PROCESSING PERFORMED

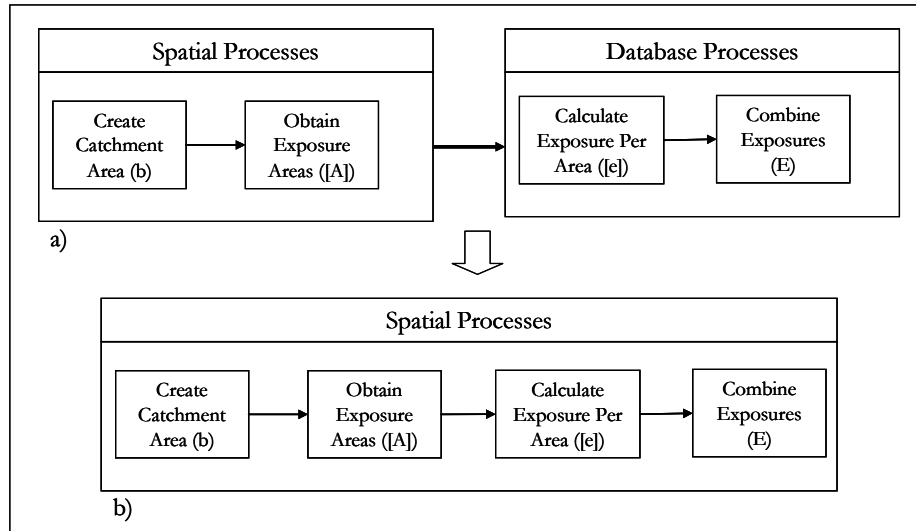
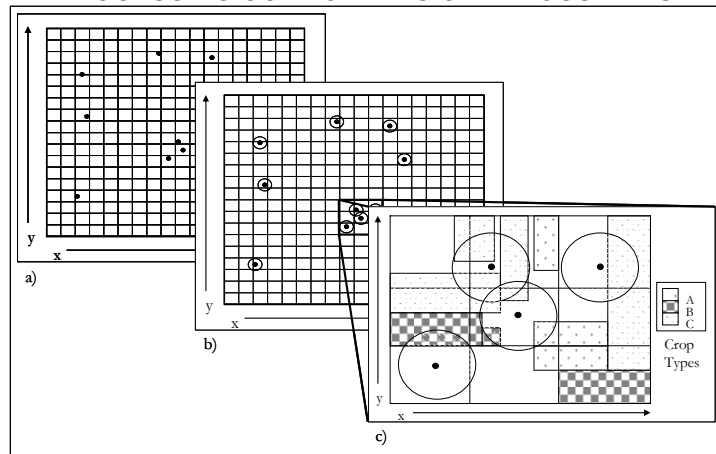


FIGURE 15
SPATIAL PROCESSING COMPONENTS OF EXPOSURE ESTIMATION



In contrast, the implementation strategy chosen for the generalized model described in this work is formulated as in Figure 14a, with processing work taking place completely within the GIS (ESRI ArcMap).

A second, related problem with the implementation described by Rull et al. (2001; 2004) was that it processed all cases together at once in each stage of the process as in Figure 16. By modeling the entire process within the GIS, we enable a per-subject processing strategy as depicted in Figure 17, where the number of subject records processed at a time can be chosen by the user.

FIGURE 16
MULTI-SUBJECT PROCESSING AT ONCE

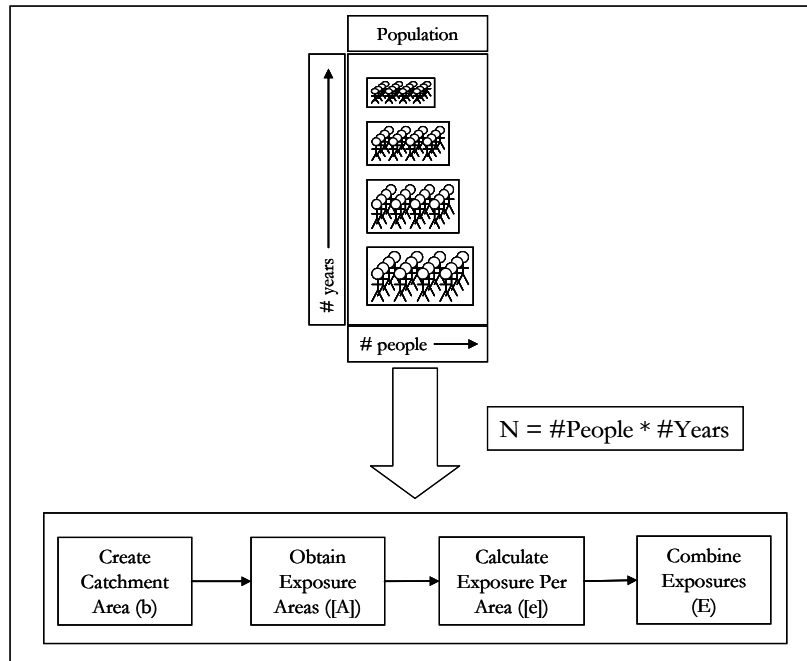
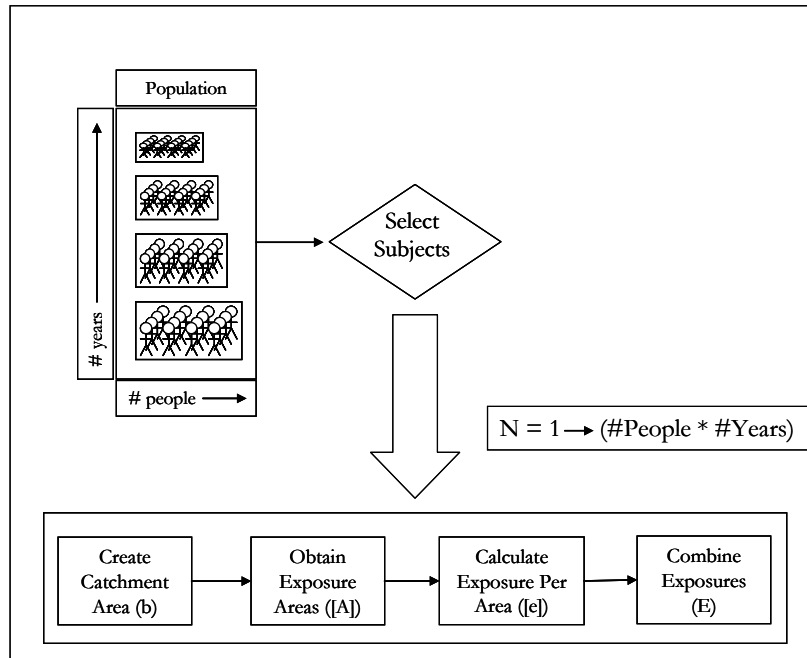


FIGURE 17
ADJUSTABLE-NUMBER-SUBJECT PROCESSING AT ONCE



INHERENT FLEXIBILITY

Perhaps the most important contribution of the exposure model we have developed is its inherent flexibility. This flexibility can be described in terms of the types of exposure data it can employ (i.e., exposure layers), the variety of spatial operations it can support (e.g., circular buffer, wind rose buffer), the mathematical operations it can use to derive exposure estimates based on different criteria (e.g., per PLSS grid cell, per crop field, tiered), and the support for multiple methods of combining per-area spatial exposures into a single exposure for the individual. By implementing the model as a series of functions that themselves take functions as parameters,

new operations can easily be incorporated into the model as they are developed. The inclusion of new data layers in the database approach would require multiple indexing across linked tables, further complicating the already bleak situation. In contrast, the per-record approach can accommodate this within the processing workflow.

In the work presented by Bell et al. (2001), Rull et al. (2001; 2003), and Nuckols et al. (2007), a circular buffer size was used. By embedding the process fully into a GIS and enabling a per-record process, we on the other hand enable the buffer size to become another variable to the process. As such, each subject can have their exposure estimated by buffers of varying sizes, e.g. at constant steps such as all buffer sizes from 10m to 10,000m at 10 m intervals. This per-record process utilizes the garbage collection of the programming language it was written in to manage the memory used and released at each iteration, and thus can be assured that the process will not be too resource intensive to either crash completely, or render the computer useless (as will occur when working in an Access database of over 2GB), thus overcoming one of the scalability limitations described by Rull et al. (2003).

The flexibility provided both by our generalized exposure model and its implementation completely within a GIS does come at a price. It is likely to generate huge volumes of data describing the exposure of each subject over long periods of time, to any number of chemicals, and at many different sizes and types of spatial catchments areas. This data, while extremely useful in epidemiologic studies, will be orders of magnitudes larger than the studies previously discussed in this paper, and needs to be consistently output from the program to some form of persistent storage. In our current implementation, the output can be controlled to write to any format for which an ODBC driver exists, e.g., Microsoft Access, Microsoft SQL Server, MySQL, or directly to a text file. This output flexibility ensures that as the amount of data produced scales to larger and larger sets, an appropriate and suitable mechanism can be obtained and utilized.

CONCLUSIONS

This paper has presented the progression of data sources and operations used to estimate environmental exposure to chemicals contained in pesticides as a test case for a more general exposure estimation methodology. A generalization of the three methodologies used in the current literature has been offered so that exposure estimates for other environmental factors can be modeled within a similar framework. From our literature search, we believe the implementation outlined in this paper to be the first published capable of encompassing the currently accepted methodologies, as well as being flexible enough to change as more accurate data sources and operations become available.

Our process, embedded completely within a GIS, offers the ability to scale graciously as the number of subjects, number of environmental factors (i.e., chemicals), and geographic variables (buffer size) are varied over time and space. Methodologies like ours will prove useful for studies wishing to perform both hypothesis generation (non-analytic studies) as well as case/control studies. The scalability provided by our method ensures that large populations of subjects can be tested along many axes of variability.

REFERENCES

- Bayer-Oglesby, L., Grize, L., Gassner, M., Takken-Sahli, K., Sennhauser, F.H., Neu, U., Schindler, C., Braun-Fahrlander, C. 2005. Decline of Ambient Air Pollution Levels and Improved Respiratory Health in Swiss Children. *Environmental Health Perspectives* 113(11), pp. 1632-1637.
- Bell, E.M., Hertz-Picciotto, I., Beaumont, J.J. 2001. A case-control study of pesticides and fetal death due to congenital anomalies. *Epidemiology* 12(2), pp. 148-56.
- Bowman, D.M.J.S., Johnston, F.H. 2005. Wildfire Smoke, Fire Management, and Human Health. *EcoHealth* 2(1), pp. 76-80.

Brody, J.G., Vorhees, D.J., Melly, S.J., Swedis, S.R., Drivas, P.J., Rudel, R.A. 2002. Using GIS and Historical Records to Reconstruct Residential Exposure to Large-Scale Pesticide Application. *Journal of Exposure Analysis and Environmental Epidemiology* 12(1), pp. 64-80.

Brody, J.G., Aschengrau, A., McKelvey, W., Rudel, R.A., Swartz, C.H., Kennedy, T. 2004. Breast Cancer Risk and Historical Exposure to Pesticides from Wide-Area Applications Assessed with GIS. *Environmental Health Perspectives* 112(8), pp. 889-897.

California Department of Pesticide Regulation. 2000. Pesticide use reporting: an overview of California's unique full reporting system. Available online at: <http://www.cdpr.ca.gov/docs/pur/purovrw/ovr52000.pdf>.

California Department of Water Resources. 2005. Land use survey. Available online at: <http://www.landwateruse.water.ca.gov/basicdata/landuse/surveys.cfm>.

Frankenberg, E., McKee, D., Thomas, D. 2005. Health Consequences of Forest Fires in Indonesia. *Demography* 42(1), pp. 109-129.

Johnston, F.H., Webby, R.J., Pilotto, L.S., Bailie, R.S., Parry, D.L., Halpin, S.J. 2006. Vegetation fires, particulate air pollution and asthma: A panel study in the Australian monsoon tropics. *International Journal of Environmental Health Research* 16(6), pp. 391-404.

Künzli, N., Kaiser, R., Medina, S., Studnicka, M., Chanel, O., Filliger, P., Herry, M., Horak, F., Puybonnieux-Textier, V., Quénel, P., Schneider, J., Seethaler, R., Vergnaud, J.C., Sommer, H. 2000. Public-health impact of outdoor and traffic-related air pollution: a European assessment. *The Lancet* 356(9232), pp. 795-801.

Marusek, J.C., Cockburn, M.G., Mills, P.K., Ritz, B.R. 2006. Control Selection and Pesticide Exposure Assessment Via GIS in Prostate Cancer Studies. *American Journal of Preventive Medicine*, 30(2S), pp. 109-116.

Microsoft Corporation. 2007. SQL Data Types [Access 2007 Developer Reference]. Available online at: <http://msdn2.microsoft.com/en-us/library/bb208866.aspx>.

Nafstad, P., Haheim, L.L., Wisloff, T., Gram, F., Oftedal, B., Holme, I., Hjermann, I., Leren, P. 2004. Urban Air Pollution and Mortality in a Cohort of Norwegian Men. *Environmental Health Perspectives* 112(5), pp. 610-616.

Nuckols, J.R., Gunier, R.B., Riggs, P., Miller, R., Reynolds, P., Ward, M.H. 2007. Linkage of the California Pesticide Use Reporting Database with Spatial Land Use Data for Exposure Assessment. *Environmental Health Perspectives* 115(1), pp. 684-689.

Reynolds, P., Hurley, S.E., Gunier, R.B., Yerabati, S., Quach, T., Hertz, A. 2005. Residential Proximity to Agricultural Pesticide Use and Incidence of Breast Cancer in California, 1988-1997. *Environmental Health Perspectives* 113(8), pp. 993-1000.

Rigel, E.G., Lebwohl, M., Rigel, A.C., Rigel, D.S. 2003. Daily UVB exposure levels in high-school students measured with digital dosimeters. *Journal of the American Academy of Dermatology* 49(6), pp. 1112-1114.

Rull, R.P., Ritz, B. 2003. Historical pesticide exposure in California using pesticide use reports and land-use surveys: an assessment of misclassification error and bias. *Environmental Health Perspectives*, 111(13), pp. 1582-1589.

Rull, R.P., Ritz, B., Krishnadasan, A., Maglinte, G. 2001. Modeling Historical Exposures from Residential Proximity to Pesticide Applications. In *Proceedings of the Twenty-First Annual ESRI User Conference*, San Diego, CA..

Rull, R.P., Ritz, B., Shaw, G.M. 2006. Neural Tube Defects and Maternal Residential Proximity to Agricultural Pesticide Applications. *American Journal of Epidemiology* 163(8), pp. 743-753.

Steinmaus, C., Lu, M., Todd, R.L., Smith, A.H. 2004. Probability Estimates for the Unique Childhood Leukemia Cluster in Fallon, Nevada, and Risks near Other US Military Aviation Facilities. *Environmental Health Perspectives* 112(6), pp. 766-772.

Tatalovich, Z., Wilson, J.P., Cockburn, M. 2006. A Comparison of Thiessen Polygon, Kriging, and Spline Models of Potential UV Exposure. *Cartography and Geographic Information Science* 33(3), pp. 217-231.

Thieden, E., Philipsen, P.A., Heydenreich, J., Wulf, H.C. 2004. UV Radiation Exposure Related to Age, Sex, Occupation, and Sun Behavior Based on Time-Stamped Personal Dosimeter Readings. *Archives of Dermatology* 140(2), pp 197-203.

Viswanathan, S., Eria, L., Diunugala, N., Johnson, J., McClean, C. 2006. An analysis of effects of San Diego wildfire on ambient air quality. *Journal of the Air & Waste Management Association* 56(1), pp. 56-67.

Ward, M.H., Nuckols, J.R., Weigel, S.J., Maxwell, S.K., Cantor, K.P., Miller, R.S. 2000. Identifying Populations Potentially Exposed to Agricultural Pesticides Using Remote Sensing and a Geographic Information System. *Environmental Health Perspectives* 108(1), pp. 5-12.

ACKNOWLEDGEMENTS

This work was supported by the National Institute of Environmental Health Sciences (Grant #5P30ES07048), the National Cancer Institute (Grant #5R03CA110846), and the Department of Defense Prostate Cancer Research Program (Grant #PC051037).