DEMOGRAPHY AND ENVIRONMENT IN GRASSLAND SETTLEMENT: USING LINKED LONGITUDINAL AND CROSS-SECTIONAL DATA TO EXPLORE HOUSEHOLD AND AGRICULTURAL SYSTEMS

KENNETH M. SYLVESTER, SUSAN HAUTANIEMI LEONARD, MYRON P. GUTMANN AND GEOFF CUNFER

INTRODUCTION

The Demography and Environment in Grassland Settlement project (DEGS) is a study of the relationship between population and environment in Kansas during its settlement and conversion from grassland to grain cultivation and rangeland.[1] The research team involved in this project (which includes professors Glenn Deane of the State University of New York at Albany and William Parton of Colorado State University, in addition to the authors of this paper) had as its goal to bring together data about farms and farm families in order to understand the core transformations in land use and family dynamics that took place during the process of settling and developing an agricultural landscape. For reasons we will explain later, the state of Kansas – located near the centre of the U.S. in a grassland ecosystem – is ideally suited for this study by virtue of its location, history and the documents that exist about it. In order to capture the environmental variability of Kansas, we are assembling a linked database of farm and family census records for twenty-five townships scattered across the state. This paper is about the process of choosing that sample, about the data we have accumulated and about the process we are undertaking to link records about families and farms through time and to attempt to find their locations in space.

The transformation of the 25 landscapes that we have chosen to represent Kansas over the 80-year period from 1860 to 1940 was part of a larger global process of grassland conversion, one that accelerated enormously after 1850.[2] In the United States access to Kansas grasslands was governed by a cadastral

History and Computing 14 (1+2) 2002 (publ. 2006), 31–60 ISSN 0 957-0144

system, first proposed by Thomas Jefferson after the revolutionary war, which surveyed land in a single continental grid west of the original 13 colonies. Beginning in Ohio, six-mile square townships, referenced to 34 north-south principal meridians, served as the basic unit of the public land survey and where numbers warranted after settlement, a territorial unit for local civil administration. The public land survey system (PLSS) further subdivided the six-mile-square township into 36 one-mile-square land parcels (sections) and each one-square-mile parcel into four half-mile-square parcels (homesteads or quarter-sections). The public land survey eventually blanketed 30 states west of Ohio with little or no correction for topography, excluding only Tennessee, Kentucky, Texas and small portions of the inter-mountain west.

The imposition of this geometrically uniform grid is sometimes viewed as only the first misstep in a series of indiscriminant uses of North America's temperate grasslands. Narratives of the historical impact of settlement often frame the discussion of institutional and human imprints in linear terms, regarding agricultural expansion as homogenizing, spatially extensive and open-ended. Before the Dust Bowl, historians often presented a progressive story, in which heroic Euro-Americans struggled against and overcame a challenging environment and hostile Native Americans.[3] They saw the linear transformation of the Great Plains as a positive outcome, one that established American communities and institutions across the continent. Ever since the 1930s, however, the association between settlement and environmental degradation has been cemented in the historical imagination.[4] These authors also described a linear process, but one following a downward trajectory, from virgin prairie to misused land to ecological collapse in the Dust Bowl. We believe that the association between settlement and misuse of land is largely asserted rather than demonstrated and that it borrows heavily from very genuine present concerns about use of technology and inorganic inputs in agriculture since the 1950s.[5] Arguably, the pace of land transformation has never matched the availability of new technology, not prior to the dust bowl in the 1930s,[6] and certainly not today, when just over 30 per cent of the U.S. Great Plains is used as cropland.[7] In contrast, we hypothesize a more iterative process: where demographic priorities shaped the transformation of landscapes, where land-use choices altered in response to monitoring of ecological processes and where transformed landscapes engendered new patterns of land stewardship and demography.[8] We motivate this research primarily as an effort to historicize long-term changes in agricultural land use in the grassland settlement of the United States. Our central objective is to understand these reciprocal processes at the level of the individual farm and family, to integrate human and natural components and to investigate whether a homogenizing, reductionist, signature was imprinted on the plains ecosystem during the settlement period.

The advantages of the study areas we have chosen in Kansas are empirical and temporal. First, the populations identified are observable during a key transformation of the grassland environment (from native grassland to cropland and rangeland and from biological to mechanized and inorganic technology), over a minimum of three human life cycles. Second, the study areas embrace the major east-west variations (vegetation, precipitation and soil texture) and to a lesser extent north-south variations (temperature) in the plains ecosystem, synoptic of wider conditions in the Great Plains and the biophysical constraints imposed on agricultural land use. Third, by following individuals and families in the community over time, our research design allows us to directly compare the decisions of households within and across ecological settings.

Kansas became a Territory in 1854 and achieved statehood in 1861. The first available federal enumeration of population and agriculture to include the state of Kansas was conducted in 1860.[9] Because the state of Kansas also conducted its own censuses, agricultural populations in the state can be observed in manuscript sources every five years from 1860 until 1940, with the exception of the missing 1890 federal manuscript census schedules. The Kansas State Board of Agriculture also enumerated agricultural land use on an annual basis beginning in 1873. Prior to 1918, farm-level manuscripts of the Kansas agricultural enumerations were systematically preserved (at the state archives in Topeka) to coincide only with federal enumerations (1860, 1870 and 1880) and state population enumerations (for 1865, 1875, 1885, 1895, 1905, 1915 and 1925). After 1918, the annual agricultural enumerations are preserved by the state archives with few exceptions and are open to the public through 1981.

SAMPLE DESIGN

Our interest in testing the effects of contextual (group-level) variables, such as precipitation and soil texture or taxonomy, led us to adopt a multilevel sampling strategy in which farms were nested within townships. Following the vocabulary of multi-level analysis, we speak of level one units, which are the smallest unit of interest (in our case, farms) and level two units, which are larger units that encompass the level one units (in our case townships). Snijders and Boskers suggest that multilevel models have important advantages in integrating the effects of context but also carry a number of assumptions that must be met:[10] most notably, that the main effects of the higher-level contexts (townships) and the main effects of the individuals (farms) be normally distributed. If this assumption is poorly approximated, results obtained in the random coefficient model may not be reliable. When the number of individuals (n_j) is moderate, say less than 100, a random coefficient model is generally preferred if the number of groups (N) is sufficiently large. Snijders and Boskers propose a 'rule of thumb' that N should be at least ten when the number of individual level (n_j) cases is small or intermediate. Recent simulation studies by Hox and Maas[11] suggests that Snijders and Boskers rule of thumb is an insufficient number. Hox and Maas found that fixed regression coefficients and variance components are estimated without bias for individual level parameters in datasets with small numbers of group level contexts (i.e., townships), but that group-level variance components are biased upwards (resulting in standard errors on group effects that are much too small). The Hox and Maas simulations show that convergence and second- (group-) level variance estimation are much improved with increasing N. Accordingly, these considerations pushed us in the direction of 25 townships rather than the minimum of 10 suggested by Snijders and Boskers. We estimated that this would yield, roughly, an average of between 1,000 to 2,500 level-one units (farms) in 25 level-two units (townships) for each time panel. After data entry was completed this turned out to be very close to the actual totals. From a total of 1,306 farms in the 1860 sample, the number of farms in each time panel rose to 2,299 by 1895 and eventually peaked in the middle of the Great Depression, with a total of 2,661 farms in the 1935 sample.

Within each township we have included all households and farms in our sample, in order to improve the likelihood of linking records over time. While statistical power is not generally compromised in multi-level models by sampling at level one, linking farms and people over time in a spatially explicit and historical study required a more comprehensive approach. Without *a priori* knowledge of the longitudinal attrition faced by a sample of historic farms, or how large local areas needed to be to capture sufficient repeated observations, the advantages of a complete universe were self-evident. The chosen design captured data for entire landscapes, not just particular farms and better integrated environmental data not typically available at spatial scales as fine as individual farms. It also held out the possibility of constructing complete universe measures, such as rates of in-migration and out-migration and using them as higher level covariates in multi-level models.

The spatial distribution of the level two units (townships within counties) in our sample was designed to capture meaningful variation in precipitation, temperature, soil quality, topography, native vegetation and crop potential across the state. The selection of counties balanced prior knowledge of environmental conditions and land use regimes with spatial coverage. The sample design had to contend with the human dimensions of county and township formation, which are the units of population information collection and yet be representative of biophysical settings that cut across those socially constructed boundaries. We chose a spatially dispersed, chessboard pattern, rather than a randomized design, in order to maximize the representation of environmental variation across the state. There were simply too many environmental variables of interest, too much space to index and too many availability issues with the historical records to limit spatial coverage or randomize the selection process.

The sample design that we have adopted effectively captures well known patterns of environmental variation in the state. Average temperatures follow a southeast to northwest gradient in Kansas, generally remaining cooler towards the northwest of the state, while average annual rainfall follows a more longitudinal pattern, increasing from west to east (see Figure 1). Precipitation is the main driver of vegetation conditions in Kansas, with the notable exception of the Flint Hills region in the east-central portion. In the western high plains, pre-settlement vegetation was dominated by hardy shortgrass species such as blue grama and buffalo grass, both very tolerant of drought and heavy grazing. In the central part of the state, as precipitation increases there is a transition to mixed grass prairies, in an ecological zone that James Malin described as the central wheat belt (see Figure 2).[12] Native Americans were farming in this area long before European contact and settlement. Francisco Vasquez de Coronado encountered horticultural villages near the Arkansas River during his expedition for the Spanish crown in 1542, as did Etienne Véniard de Bourgmont while acting as an emissary of the French crown in 1722.[13] Further east, in a zone bounded to the north by the Kansas River and to the south by Oklahoma's state boundary, a 70 mile-wide corridor of bluestem pasture punctuates the flatland prairie. Texas cattle ranchers once drove their herds to be fattened in the bluestem tallgrasses and the region, known as the Flint Hills, remains predominantly in grassland cover today because of the hilly terrain. In the eastern fifth of the state, mixed farming has predominated in a region that was largely tallgrass prairie and forest mosaic prior to settlement. Along a northern tier of counties near the border with Nebraska is a zone that became known for its capacity to support corn and soybean cultivation.

There are two types of townships in Kansas (as noted above): the six-mile-square township of the Public Land Survey System and the political townships created by counties to administer local affairs. In hindsight, we began with a somewhat naïve notion that the PLSS would be followed in the creation of minor civil divisions across Kansas. To some extent this was true of the central part of the state and along its northern border. Civil townships that were neatly organized within the coordinates of the PLSS made our task of indexing space much easier. Typically these minor civil divisions were organized during the dramatic period of post-Civil War expansion, when the formation of townships conforming to the grid simplified the land development process for county commissioners. However, in the west and east, the design of townships did not follow the grid. In the eastern third of the state local initiative and topography often triumphed over the geodetic perfection and county commissioners created units of varying shape and size, even some smaller than the survey township. By contrast, in the western third of the state population densities remained so low that county commissioners were obliged to create larger territorial units. Legislation passed in 1868 tried to limit the formation of new townships, directing that county commissioners not organize new townships until local populations

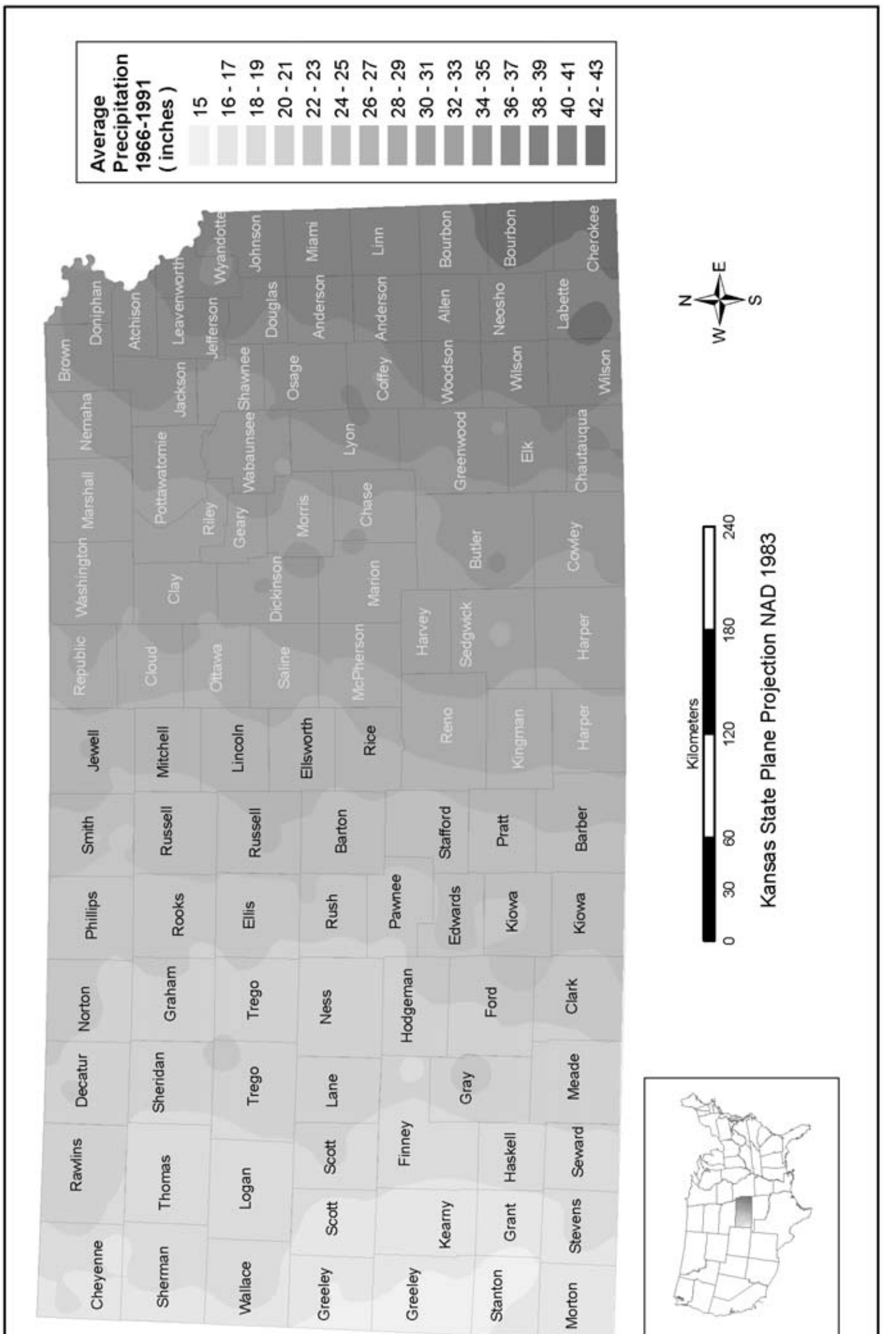

Figure 1. Average annual precipitation in Kansas, 1961–1990.
Source: Natural Resources Conservation Service, 1998.
http://www.ncgc.nrcs.usda.gov/products/datasets/climate/data/precipitation-state/ks.html

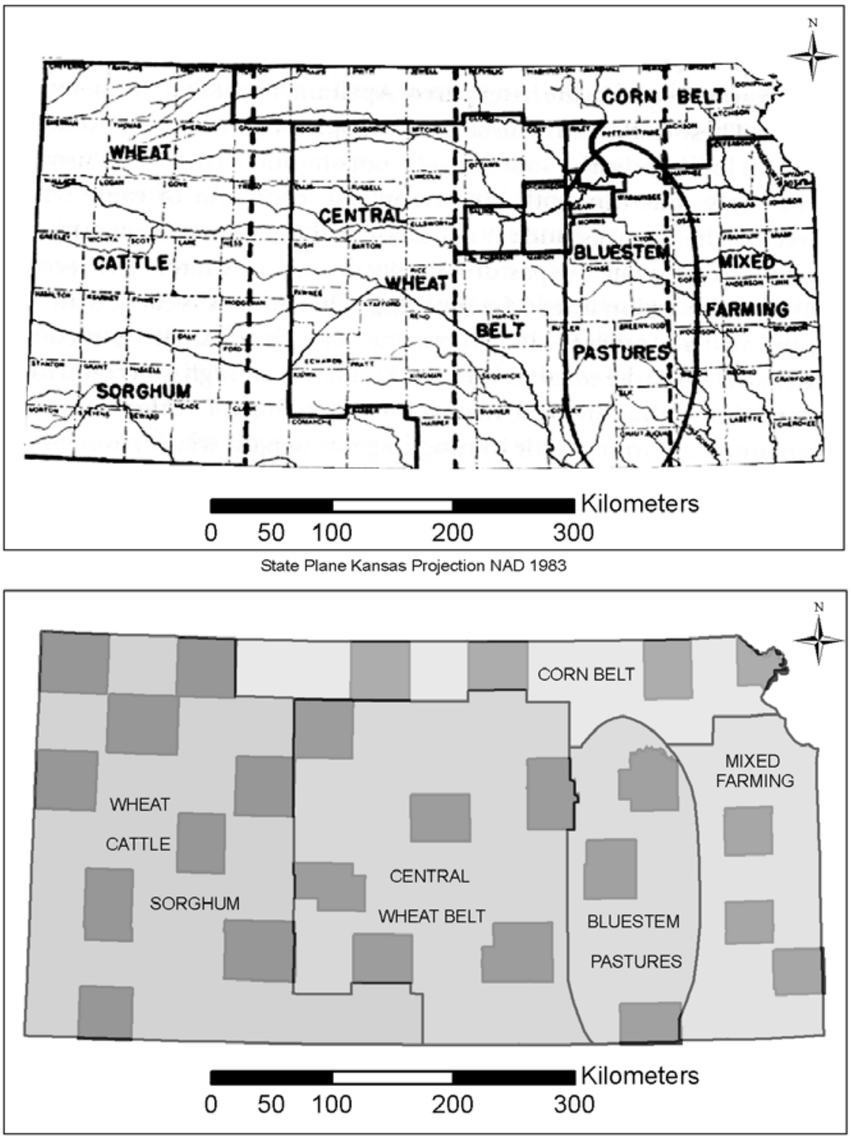

Figure 2. Target counties and Malin's land use zones.
Source: original map from James C. Malin: *Winter wheat in the golden belt of Kansas: a study in adaption to subhumid geographical environment* (Lawrence, KS: University of Kansas, Press 1944), preface.

reached two hundred souls, a petition was made by 50 electors to subdivide the county and the requested change did not create a township smaller than 30 square miles in area.[14]

The changes that were made in the period between 1874 and 1888 are well documented in the volumes of the Kansas State Board of Agriculture's *Biennial reports*, where a detailed series of color township maps were reproduced every two years. The maps illustrate where minor civil divisions deviated from our hoped-for symmetry and conveniently overlay these boundaries on the co-ordinates of the PLSS grid. The State Board of Agriculture's maps made the process of sample selection easier by affording an opportunity to digitize and geo-reference boundary maps for use in township selection (see Figure 3).[15] As the selection process moved forward, the digitized maps proved invaluable for assessing how well 'candidate' townships met our selection criteria. After considering several alternatives, we settled on three key exclusion rules and two inclusion rules. Townships should be excluded, we reasoned, if the land in the township were at any time during the study period: (1) a functioning military reserve or base; (2) an Indian reserve; or (3) if data for the township was missing or of poor quality. Townships should be included only if: (1) the land in the township remained predominantly agricultural during the study period; and (2) the township in question had the highest rural population density in the county in 1910. Our main hypothesis was that after a period of trial and error, effectively over by the time of the 1910 census, farmers and farm families would gravitate to the most promising landscapes within each county. Therefore, by choosing the most densely settled rural townships, we would be selecting from among the most intensively managed or human dominated agricultural landscapes within each of the 25 target counties.

Even with reasonable prior knowledge, township selection proved to be a more iterative process than we expected. Digital map files of historical boundaries of minor civil divisions were not available. We began by making an investment of time to digitize historical townships maps for 25 target counties and attribute them with population data, drawn from the Kansas State Board of Agriculture's *Biennial reports*. This involved scanning the historical township maps and geo-referencing the image files to recognizable spatial coordinates (see Figure 3). We used digital map files, in shapefile format, of the public land survey system in each county.[16] We then overlaid the same public land survey files on the geo-referenced images and used the information in the image files to merge the one-mile-square polygons, or sections, in the shape files until we had created new polygons that conformed to the civil townships in the counties. This effort allowed us to perform some thematic mapping of population data – to visualize population density distributions against the land use zones we had sampled for in our target counties (see Figure 4). The mapping exercise also paid further dividends by forcing us to think through township boundary changes and to

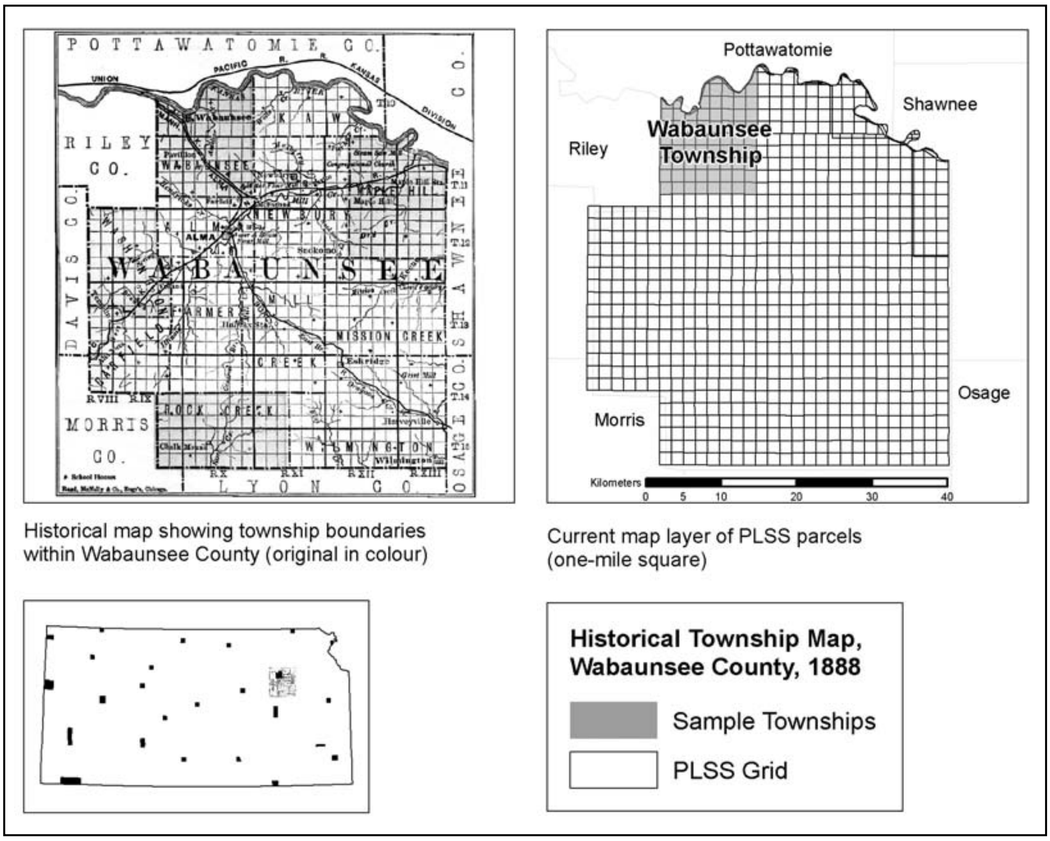

Figure 3. Public Land Survey Grid overlaid on geo-referenced historical map, Wabaunsee County, 1888.
Source: Kansas State Board of Agriculture, *Biennial reports*, 1887–88.

relate our population collection units to ancillary environmental data, such as current land use mapping or soils data.

Before we finalized the township selections, however, we had to deal with an unexpected revision of the sample. About a month or so into the process, with research staff scanning and geo-referencing township maps from our original selection list, we discovered that a key set of evidence was missing from the archival record. The agricultural schedules from the 1880 federal census were missing for Kansas counties whose names fell alphabetically within a range of the letters G and J (Geary, Graham, Gray, Greeley, Greenwood, Hamilton, Harper, Harvey, Haskell, Hodgeman, Jackson, Jefferson, Jewell and Johnson). We contacted the National Archives and the Kansas State Historical Society and neither was aware of a gap in the census records, but both reluctantly confirmed

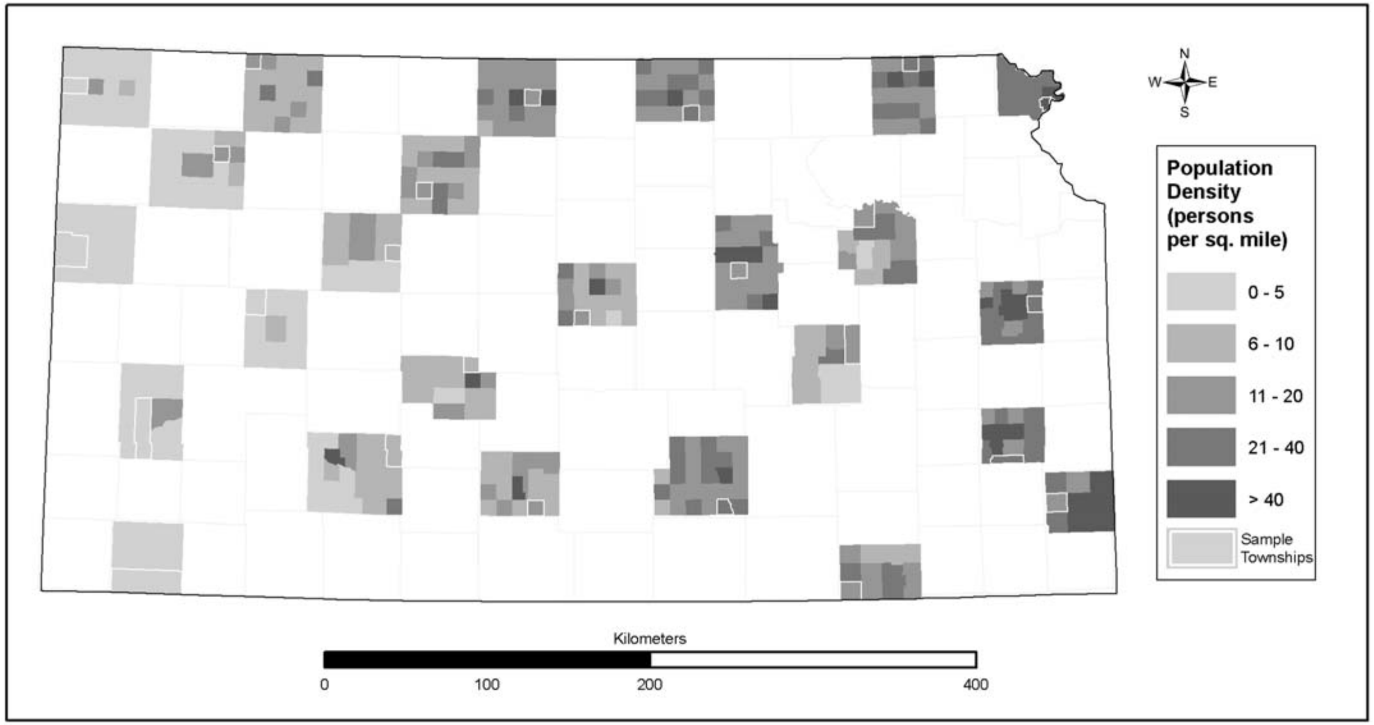

Figure 4. Sample townships and population densities, by township, within target counties, in 1910.
Source: Kansas State Board of Agriculture, *Biennial report*, 1910.

our fears that counties beginning with these letters were in fact missing. No matter how the manuscript records were lost, there was no way around the omission. We looked at our county maps again to see if there were counties that didn't start with a letter between F and J that could preserve our spatial design and found several candidates for replacements. We also learned from published sources that Comanche, Finney, Saline and Seward counties were not enumerated in 1895.[17] These candidates had to be dropped from the sample. Fortunately, in the end, there were enough target counties.

In the final iteration of township selection, we also tried to minimize inclusion of townships that experienced substantial boundary changes. Ultimately, we excluded six townships that met most of our selection criteria but whose boundary changes posed significant data quality issues.[18] In choosing replacements, our aim was to select townships that were smaller subsets of larger predecessors and which did not differ greatly in terms of population density from other rural townships in the target counties. As Figure 4 illustrates, the latter was a straightforward tradeoff because population densities varied far less within counties than they did across the state. This still left four alternative townships with boundary changes. In Stevens County, for example, commissioners divided the county into three equally sized north-south rectangular units and then, after actual settlement made these divisions unworkable, split the county into three equally sized east-west townships. But in the case of each replacement township, the successor township was a smaller subset of the larger predecessor, obliging us to gather population information from a larger area at an earlier time, but nevertheless allowing us to index a consistent political space going forward in time.[19] Table 1 lists our final selection of 25 civil townships, approximate location on the PLSS grid, population densities at key time points, area in square miles and an indication of changes to township boundaries. Ultimately, 14 townships were bounded by the six by six mile township of the PLSS grid, one township that was only half that size and 10 townships occupied a much larger area.

Beyond helping to select the sample townships, the effort to digitize minor civil boundaries also allowed us to begin the work of generating variables from environmental data, to serve as township level covariates in multilevel models. Historical monthly weather data (precipitation, minimum and maximum temperature) from the PRISM project were summarized to township boundaries from 1895 to 1940.[20] Soils data from Natural Resources Conservation Service's new soils series, NASIS (National Soil Information Series), were also summarized to fit our township boundaries.[21] The NASIS variables generalized to our township boundaries included capability class, erosion class, soil runoff potential, flooding frequency, hydrologic group, soil erodibility factor, structure size, wildlife rating, wind erodibility, farmland classification, taxonomic soil order and texture class.

Table 1. Townships selected.

County	*County FIPS*	*Township*	*PLSS township*	*PLSS range*	*PLSS meridian*	*Population density per sq. mile 1888*	*Population density per sq. mile 1910*	*Population density per sq. mile 1940*	*Village population 1888*	*Village population 1910*	*Village population 1940*	*Boundary changes*	*Area (square miles)*
Allen	001	Cottage Grove	26	18	6	21.6	22.4	16.4	0	0	0	**Y**	47.39
Chase	017	Toledo	18	9	6	6.9	10.2	7.8	0	0	0	**Y**	135.60
Chautauqua	019	Harrison	34	9	6	15.1	13.5	8.9	0	0	0	N	54.26
Cheyenne	023	Cherry Creek	3	42	6	4.3	4.6	5.2	0	0	0	N	59.44
Crawford	037	Grant	29	21	6	21.0	17.1	14.4	0	0	0	N	55.37
Decatur	039	Finley	1	30	6	0.0	12.1	7.9	0	0	0	N	35.14
Dickinson	041	Newbern	14	1	6	18.4	18.9	11.9	0	0	0	N	35.89
Doniphan	043	Marion	3	21	6	30.3	42.8	31.2	0	0	0	N	16.98
Elsworth	053	Green Garden	17	9	6	15.9	14.7	15.3	241	0	0	N	35.89
Ford	057	Wheatland	25	21	6	7.7	6.5	6.0	0	0	0	N	73.06
Franklin	059	Peoria	16	20	6	30.0	26.1	17.0	0	0	0	**Y**	35.07
Kearny	093	Hartland	23	37	6	nd	2.6	1.1	nd	0	0	N	137.64
Lane	101	Cheyenne	16	30	6	5.0	3.9	5.0	0	0	0	N	77.54
Nemaha	131	Washington	1	13	6	22.5	26.4	20.7	0	0	234	N	35.32
Pawnee	145	Walnut	20	16	6	8.0	8.4	6.0	0	0	0	N	35.46
Pratt	151	Paxon	29	12	6	11.0	15.6	10.0	0	0	120	N	35.76
Republic	157	Grant	4	2	6	21.0	21.0	15.1	0	0	70	N	36.03
Rooks	163	Logan	9	19	6	13.2	15.8	12.6	0	0	0	N	35.51
Sedgwick	173	Salem	29	1	6	23.9	20.8	19.2	0	0	0	N	31.91
Smith	183	Blaine	3	12	6	16.4	16.8	10.6	0	0	0	N	35.44
Stevens	189	Voorhees	35	39	6	nd	5.0	2.3	nd	0	0	**Y**	238.02
Thomas	193	Lacey	7	32	6	5.7	13.0	5.9	0	0	0	N	35.98
Trego	195	Glencoe	13	21	6	nd	6.3	5.7	nd	0	0	N	35.40
Wabaunsee	197	Wabaunsee	10	10	6	16.1	15.4	10.9	0	0	0	N	65.88
Wallace	199	Weskan	12	43	6	nd	2.7	2.5	nd	0	0	N	153.25

Bold type = townships with boundary changes c. 1888–1940.
nd = no data reported.
Source: Kansas State Board of Agriculture, *Biennial reports*, 1887–88, 1909–10, 1939–40.

It is fascinating to compare the population densities observed in the sample townships in 1910 with current satellite imagery of land cover in the 1990s. Figure 5 displays our sample township boundaries superimposed on recent land cover imagery. Developed as part of a USGS program, the National Land Cover Data (NLCD) are derived from 30 meter Thematic Mapper satellite imagery, classified with a modified version of the Anderson Level II system.[22] The 21 land cover classes reported in the 1992 NLCD indicate that land uses in the study townships remain very close to the predominant land uses characterized by Malin in the 1930s. In fact, zonal sums of the NLCD land cover classes within each our townships indicate that small grains remain the dominant land cover in a land use zone Malin styled as the Central Wheat Belt, that grassland predominates in the western Wheat, Cattle and Sorghum zone and in the Bluestem Pasture region, that improved pasture or hay is the dominant land cover in the Mixed Farming zone and that row crops (corn, soybeans, sorghum) represent a majority of land cover only in the two most easterly townships in the northeastern Corn Belt zone. Our township sample demonstrably reflects these enduring land use regimes. Our challenge is to understand when the sensitivity to environmental conditions now visible far above the earth was manifest in households responsible for the stewardship of the grassland.

POPULATION AND FARM DATA

We employed research assistants to key enter all the information items in the censuses for all persons and farms at each time point within the chosen townships.[23] The Kansas population census asked fewer questions than did the federal census, but added questions on military service, apprenticeship and migration. Table 2a shows the content of our 14 population data panels, illustrating the most important and consistent questions across the federal and state population censuses from 1860 to 1930. The darker shading indicates information from the federal census, lighter shading from the Kansas state census and blank cells indicate no data was collected for that question in that year. Similarly, Table 2b shows the information gathered from 14 agricultural data panels, grouped by the major categories of information, available from 1860 to 1940. Because settlement proceeded across Kansas from east to west over a period of several decades, data panels are not uniform until 1895 (see Table 2c). In total, we have population and agricultural census data for nearly 300 township/year combinations. We have records for over 173,500 individuals in nearly 44,000 households and for almost 30,000 farms. The number of farms in our townships peaked in 1935 at around 2,660. Population peaked at nearly 16,000 in 1910, varying from a total of 233 persons enumerated in Glencoe Township, Trego County, in the western portion of the state, to 1,053 in Wabaunsee Township, Wabaunsee County, in the east. Population density was

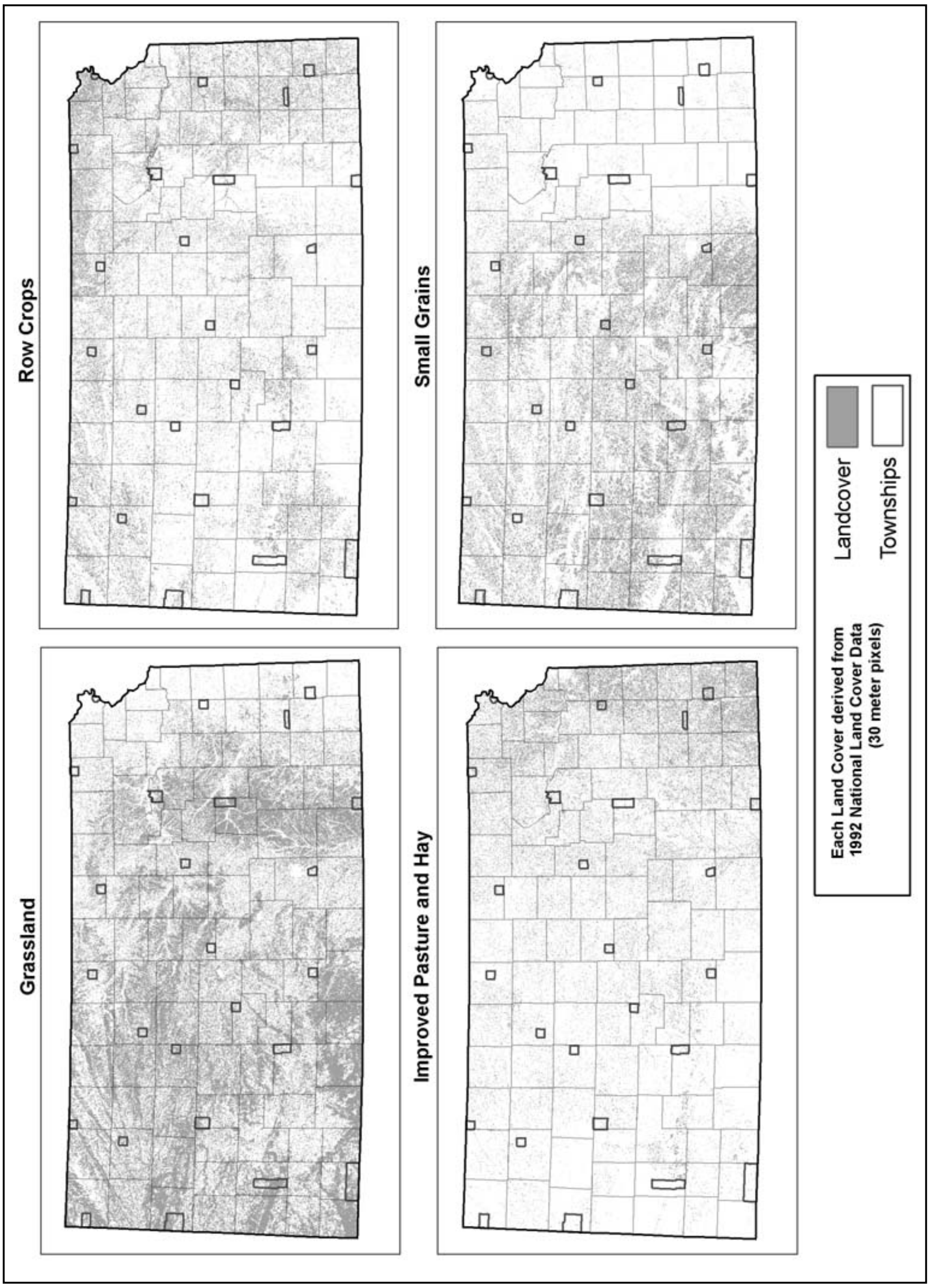

Figure 5. Sample townships and land cover in 1992.
Source: Derived from USGS National Land Cover Data, 1992, http://landcover.usgs.gov/natllandcover.asp

Table 2a. Population census questions.

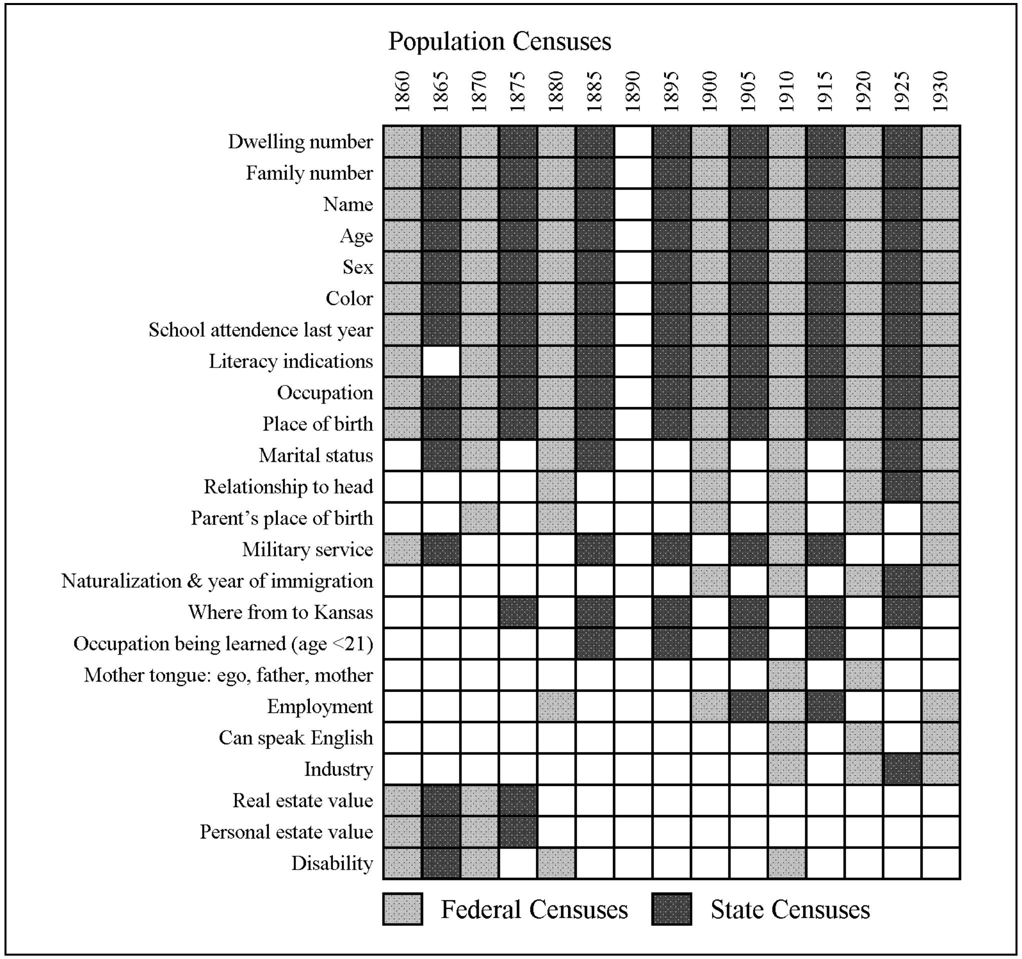

Table 2b. Agricultural census questions.

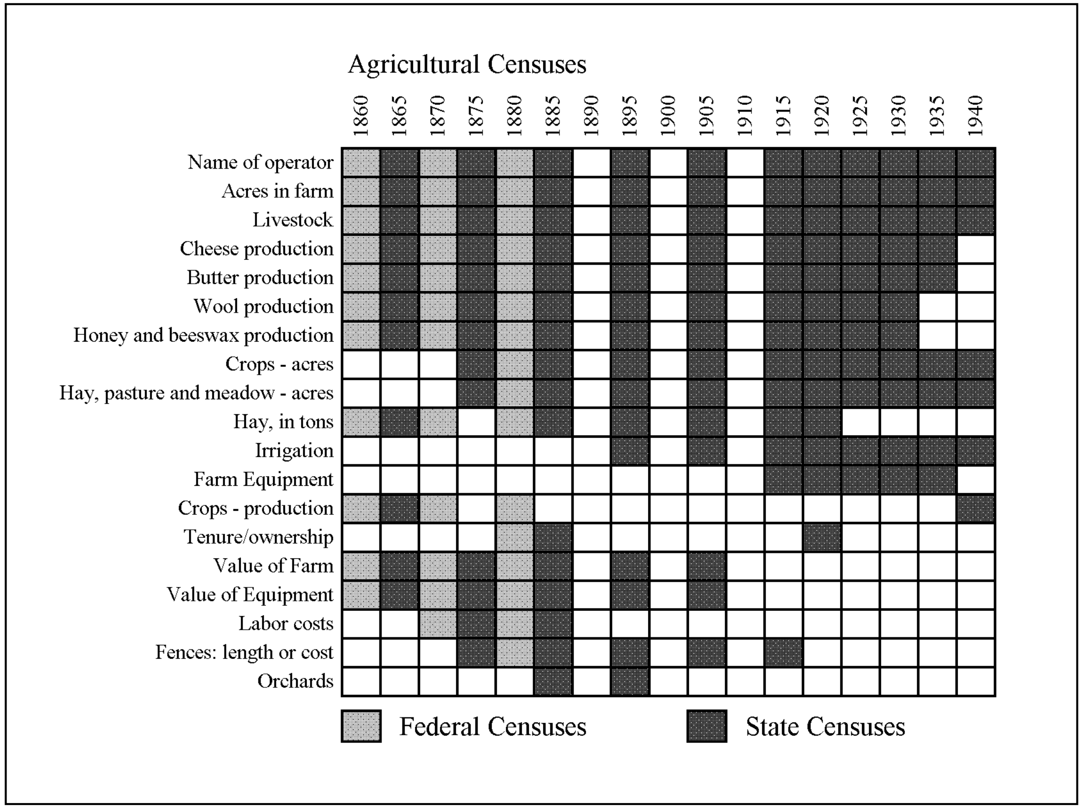

Table 2c. Township census availability.

County	Township	Total P	Total A
Allen	Cottage Grove	14	14
Chase	Toledo	14	13
Chautauqua	Harrison	13	12
Cheyenne	Cherry Creek	10	10
Crawford	Grant	13	13
Decatur	Finley	10	10
Dickinson	Newbern	14	14
Doniphan	Marion	14	14
Ellsworth	Green Garden	12	12
Ford	Wheatland	11	11
Franklin	Peoria	14	14
Kearny	Hartland	8	8
Lane	Cheyenne	9	9
Nemaha	Washington	14	14
Pawnee	Walnut	12	12
Pratt	Paxton	10	10
Republic	Grant	12	12
Rooks	Logan	10	10
Sedgwick	Salem	12	12
Smith	Blaine	12	12
Stevens	Vorhees	9	8
Thomas	Lacey	11	11
Trego	Glencoe	10	10
Wabaunsee	Wabaunsee	14	14
Wallace	Weskan	9	9
Grand Total		291	288

Federal Censuses

State Censuses

also quite varied across sample townships in 1910, ranging from 2.7 persons per square mile in Weskan Township, Wallace county, on the state's western boundary with Colorado, to 42.8 persons per square mile in Marion Township, Doniphan County along the Missouri River in the northeast corner of the state.

DATA LINKAGE

The heart of the project is linking information from the population censuses to agricultural censuses taken in the same year (horizontal linkage) and then following individuals and families in the population and agricultural censuses across time (vertical linkage). While a variety of computerized techniques have been developed for the linkage of historical records, based on computer algorithms and maximum likelihood concepts and incorporating varying information sets, we favor a combination of computerized and interactive methods.[24]

The Project has four major linkage tasks: two horizontal and two vertical. The *first* is to link farmers in the population census to farmers in the agricultural census. This is the most straightforward linkage task but also the most heavily dependent on names for matching. Last name, first name or initial and middle name or initial are available from both sets of schedules. Occupation can also be used in linking since persons listed on the agricultural schedules are mostly farmers. Individuals listed in the agricultural censuses are also much more likely to be adults and to be male and these propensities can be exploited in record linkage. Linkage success is also enhanced because these populations are relatively small and the universe of possible matches is limited, with exception of some redundancy within large families. Of course, this advantage will be offset by migration during periods of high population turnover. Perhaps most importantly, the agricultural schedules were often enumerated in the same order as the population schedules. A *second* slightly different horizontal task (for the years 1920, 1930 and 1940) is to link federal population schedules to Kansas agricultural schedules. This link is facilitated by a separate schedule in the Kansas Department of Agriculture Census which lists heads of household and the number of persons in the household. By linking the two Kansas schedules, we can see how many members were in a farmer's household and use that to assess the linkage with federal population census households. The *third* task – vertical linkage between one population census and the next – is based on matching information sets for individuals and also uses age, sex and place of birth. Linkage is based on matching information sets for individuals. Household composition, the names of household members and marital status may also be used with appropriate caution, recognizing that these items are subject to unseen changes across the time periods being linked. While more information is available for linkage across time using population records, this linkage is the most affected by death and by in and out-migration. The *fourth* task is to link farmers over time across agricultural censuses. Linkage here relies on name, place and farm size. This task is necessary to track farmers who reside outside the study townships but are listed in the agricultural schedules because their farm properties fall within the township. This last task will provide checks on the identity of farmers identified through the combination of horizontal and vertical linkage.

The data tables are managed as a series of Oracle tables in a relational database, with separate tables for each township-time-type combination and a minimum of redundant information across tables. Each person, family, household and farm at each time point has a unique identifier. Tables can be joined, queried, assembled and exported using conventional SQL commands on the basis of a linkage table and key fields in each table. (Alternatively, tables can easily be exported as text files and manipulated with other software.) Linkages can be compared across time to trace individuals, families or farms across more than two time periods.

A user-interface developed in Java by Steve Burling, our programmer at ICPSR (see Figure 6a), allows operators to retrieve, display and compare data tables stored in Oracle. Links are made by applying a hierarchy of linkage algorithms in multiple passes, with the likelihood of the algorithm producing acceptable matches used to determine the sequential order in which linkage criteria are applied. All linkage information is then stored in a in a separate relational table in Oracle, rather than appended to the original data tables. Generally, matches are retained if they meet the highest likelihood of being a true match, that is, the criteria in the strictest algorithm. At each lower level of likelihood, however, all possible matches generated by the algorithm being invoked are displayed and available for linkage, providing a check on the assumption that the strictest criteria provides the best matches. Highest likelihood criteria require the most demanding match on all items, middle-level criteria employ a type of name 'stem' matching and lowest level criteria with minimal item congruence are used to subset the data files into minimally qualified sets, which are displayed as suggestions for further sorting and interactive consideration by the operator. Generation of stems for name matching, created by removing duplicate consonants and all vowels, helps to overcome difficulties in matching ethnic names and overcomes some of the uncertainty in interpreting vowels in the handwritten manuscript records. The algorithms and the order in which they are employed are displayed in Table 3.

Figure 6a shows the results of a match using algorithm 1, the strictest. All persons on the left in an 1895 population table for Finley Township in Decatur County are displayed. In the right-hand table, the person selected by the algorithm from the 1900 population table for Finley township and all members of the dwelling in which he was enumerated are displayed. Figure 6b shows the results of algorithm 3, which employs the name stems. Essentially, all persons

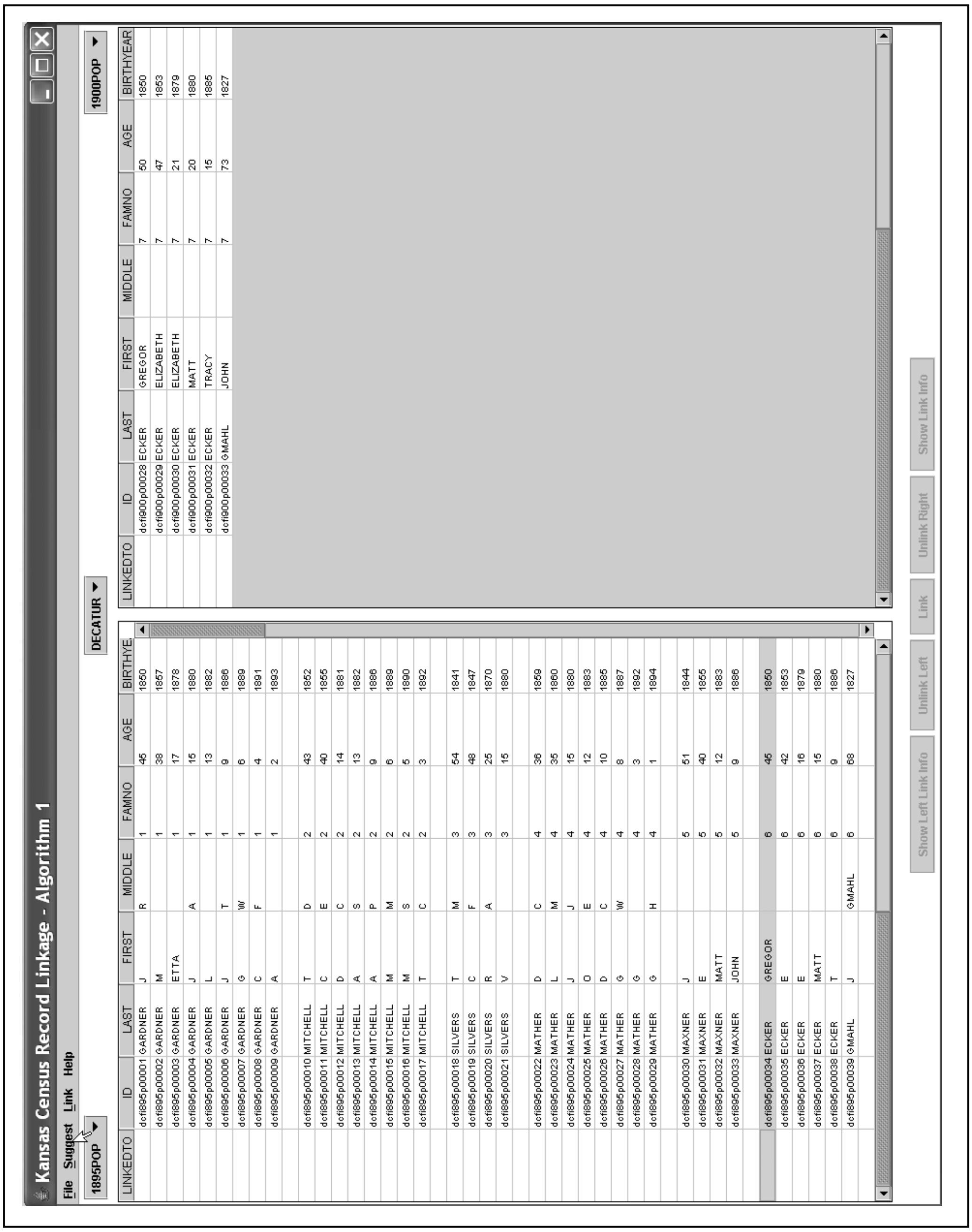

Figure 6a. Algorithm 1 the strictest link.

Table 3. Types of linkage algorithm.

Algorithm number	*Type of match*	*Name*		*Type of match*	*Name*		*Type of match*	*Name*	*Pop to pop only*	
1	exact	first	+	ignore	middle	+	exact	last	birthyear +/– 5	
2	initial	first	+	ignore	middle	+	exact	last	birthyear +/– 5	sex
3	exact	first	+	ignore	middle	+	hi score	last	birthyear +/– 5	
4	exact	first	+	ignore	middle	+	initial	last	birthyear +/– 2	
5	hi score	first	+	ignore	middle	+	hi score w/o init	last	birthyear +/– 2	sex
6	initial	first	+	ignore	middle	+	initial	last	birthyear +/– 2	sex
7	exact	first	=	exact	middle	+	exact	last	birthyear +/– 5	
8	hi score	first	=	hi score	middle	+	hi score	last	birthyear +/– 2	sex
9	initial	first	=	initial	middle	+	exact	last	birthyear +/– 2	sex
10	ignore	first	+	ignore	middle	+	exact	last	birthyear +/– 2	sex
11	exact	first	+	exact	middle	+	ignore	last	birthyear +/– 2	sex

Note: For linkage from agricultural records to population records, algorithm 1 would read 'the first name from the agricultural record exactly matches the first name from the population record *and* the last name from the agricultural record exactly matches the last name from the population record'. Algorithm 3 would read 'the first name from the agricultural record exactly matches the first name from the population record *and* comparison of the collapsed last name from the agricultural record to the collapsed last name from the population record generated a score of above 50'. For linking population records to populations records, algorithm 1 would read, 'the first name from the earlier population record exactly matches the first name from the later population record *and* the last name from the earlier population record exactly matches the last name from the second population record *and* the estimated birth year of the earlier population record is within 5 years plus or minus of the birth year from the later population record'. Birth year is estimated by subtracting age at census from census year.

Algorithms 7, 8 and 9 compare the first name from the agricultural (or earlier population record) to the middle name from the later population record. In our data, there are many instances of changes in the order of given names across records.

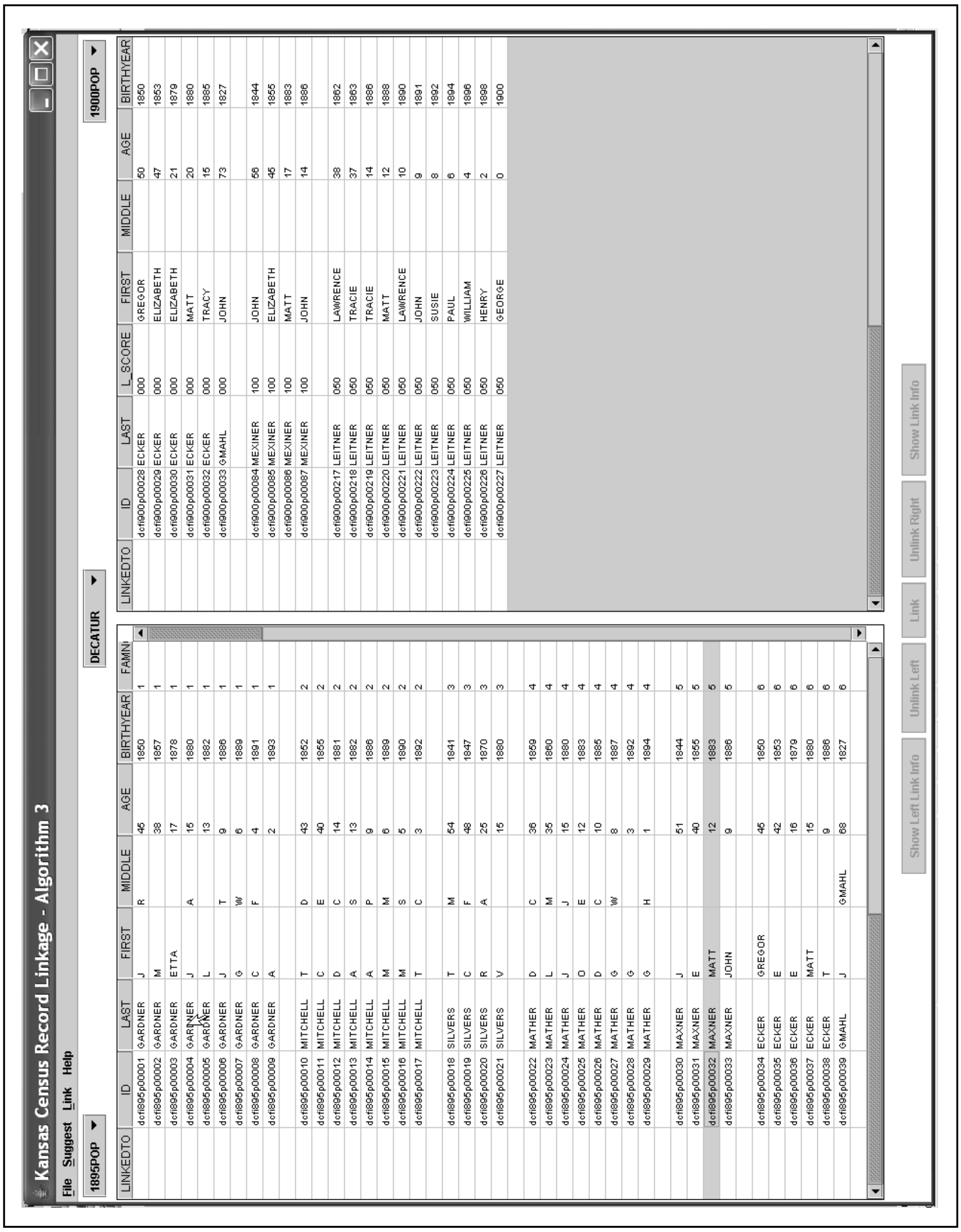

Figure 6b. Stem matching.

who meet the non-stem requirements of algorithm 3 are displayed in the right-hand table, along with a score reflecting the degree of congruence between the two stems. Records are then sorted by score and matches with a score of 50 or better may be linked. To overcome the limitation of algorithms based on initial matching, the fifth algorithm to be used employs the name stem matching ignoring the initial letter of the name. It is also possible to see all records from each of the tables and to multiply sort columns within the tables. Matches made in that mode are identified as algorithm 0.

Once an acceptable match has been identified and chosen by the operator, a 'make-link' text box appears (Figure 6c). This box contains all the information that will be stored in the link table, once the link is accepted. The unique identification numbers of the linked records, the algorithm which generated the match and whether the birthplaces and occupations match are filled in by the program. Birthplace and occupation matching are used to evaluate competing links between population tables. The make-link box also lets the operator use information about other family members, available because all members of the dwelling are displayed in the left-hand table. In the case of competing links of children, the operator can use the first names of mother and father to evaluate which is the true link; checking those boxes preserves this information about the operator's decision. Once the index person suggested by the algorithm has been linked, all other appropriate members of the dwelling can also be linked. This is indicated by checking the 'family link' button. The link table preserves the algorithm on which the index person was selected. Algorithm designations are generated post-hoc for individuals who are family-linked or linked at algorithm 0. Finally, the operator may make a comment on the record, for instance, if competing links are evaluated on criteria other than place of birth, occupation, or parent's name (e.g., farm size in the case of agricultural tables). In practice, we have seen few competing matches. Due to the small size of most of the townships, sub-setting by age and sex results in very small pools of potential links which are clearly distinguishable with very gross nominal information. When more than one match could be a link, household information clearly eliminates most matches.

We expected that most individuals on the agricultural schedules would match to the population schedules but fewer individuals would link across time: likelihood confirmed thus far by our experience working with data from Decatur and Chase counties. Linkage of population censuses in the settlement period reveals high population turnover. Of 104 households in Finley township, in Decatur county, in 1880, only seven individuals and an additional five family names remained in the community at the time of the 1900 census. This is due mainly to population turnover, but also reflects a reduction in the size of the enumeration district in 1900, when Finley was created out a larger township and the lack of intervening information between 1880 and 1900. Nevertheless, we

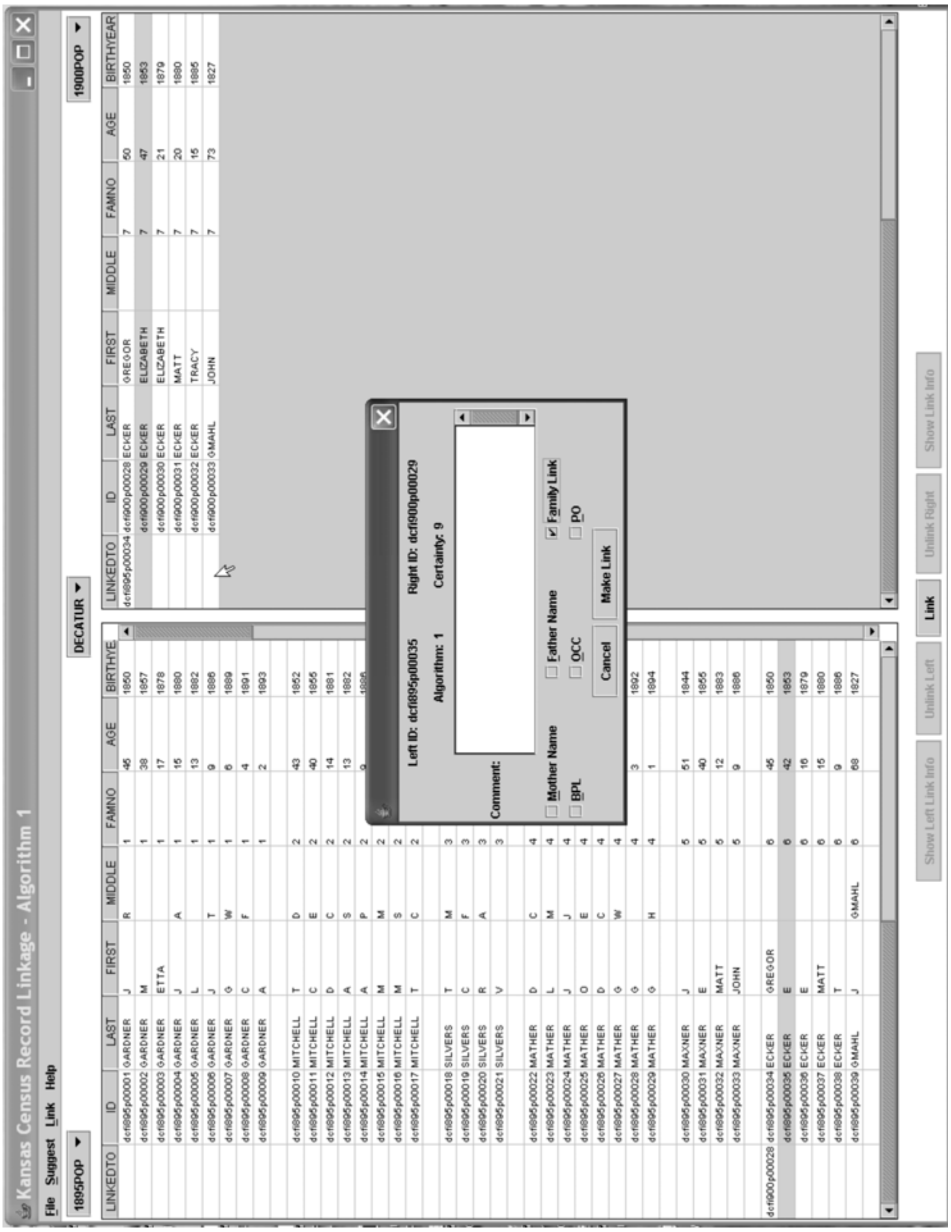

Figure 6c. Screen shot of make link box.

know from other historical studies of the phenomenon, including two based on Kansas data, that population turnover in new rural areas should decrease as agricultural settlement matures.[25] The Finley township data initially reflect these broad trends. Of 65 households in the 1900 census of Finley township, 54 per cent were found in the 1910 census (35), plus an additional 8 family names. Of 90 households in the 1910 census, 61 per cent were found in the 1920 census (55/90). Half the households in the 1925 Kansas population census were also present in the 1920 federal census (52/101). While only two household heads in the 1920 census could be traced back to the 1880 census, several family lineages are present in the community throughout the 40-year time span. Nearly one-third of the households in 1900 had at least one member present in Finley Township twenty years later (21 of 65). Of the households in 1925, 25 had arrived in town as recently as 1920, 12 were in the 1910 census and 15 were represented in the census of 1900. Horizontal linkage in Decatur County surpassed our high expectations. Overall the agricultural censuses 95 per cent of the farmers listed were linked to the population schedule. Linkage from Kansas agricultural schedules to federal population schedules were the lowest, at 87 per cent in 1930 and 91 per cent in 1920 (9 and 6 unlinked records, respectively). Four of the unlinked farmers were enumerated in nearby towns and 2 were enumerated in the population census 5 years later. Linkage from Kansas agricultural schedules to Kansas population schedules was nearly perfect, ranging from 96 per cent to 100 per cent, with at most 3 farm records not linked in any given year.

SPATIAL RESOLUTION OF FARMS AND FARM FAMILIES

Finally, since only the township and post office of the farm households are listed in the census records, we have begun to experiment with various methods of placing people and farms in specific landscapes. The most cost effective method of establishing the location and spatial extent of farms in relationship to the ownership record is to digitize parcel maps produced during the period by land companies and gazetteers. These commercial maps typically provided detail on parcel ownership, displaying names and subdividing lots directly on maps of places like Finley Township in Decatur County. For instance, in the plat maps reproduced in Figure 7, 49 of the 89 owners listed on the 1920 map of Finley township could be cross-referenced to the 1920 agricultural census. We have searched archives and digitized all of the parcel maps we can find. For some counties only one date is available. For most we have land ownership maps for two or three years and for some as many as five. Nemaha County, for example, has parcel maps available for 1887, 1908, 1922 and 1949. We scanned each parcel map and geo-rectified it to the Public Land Survey System boundaries in a GIS. To digitize farm boundaries we overlay the PLSS shapefile of quarter sections on the scanned map images. Since many Kansas farms began as quarter

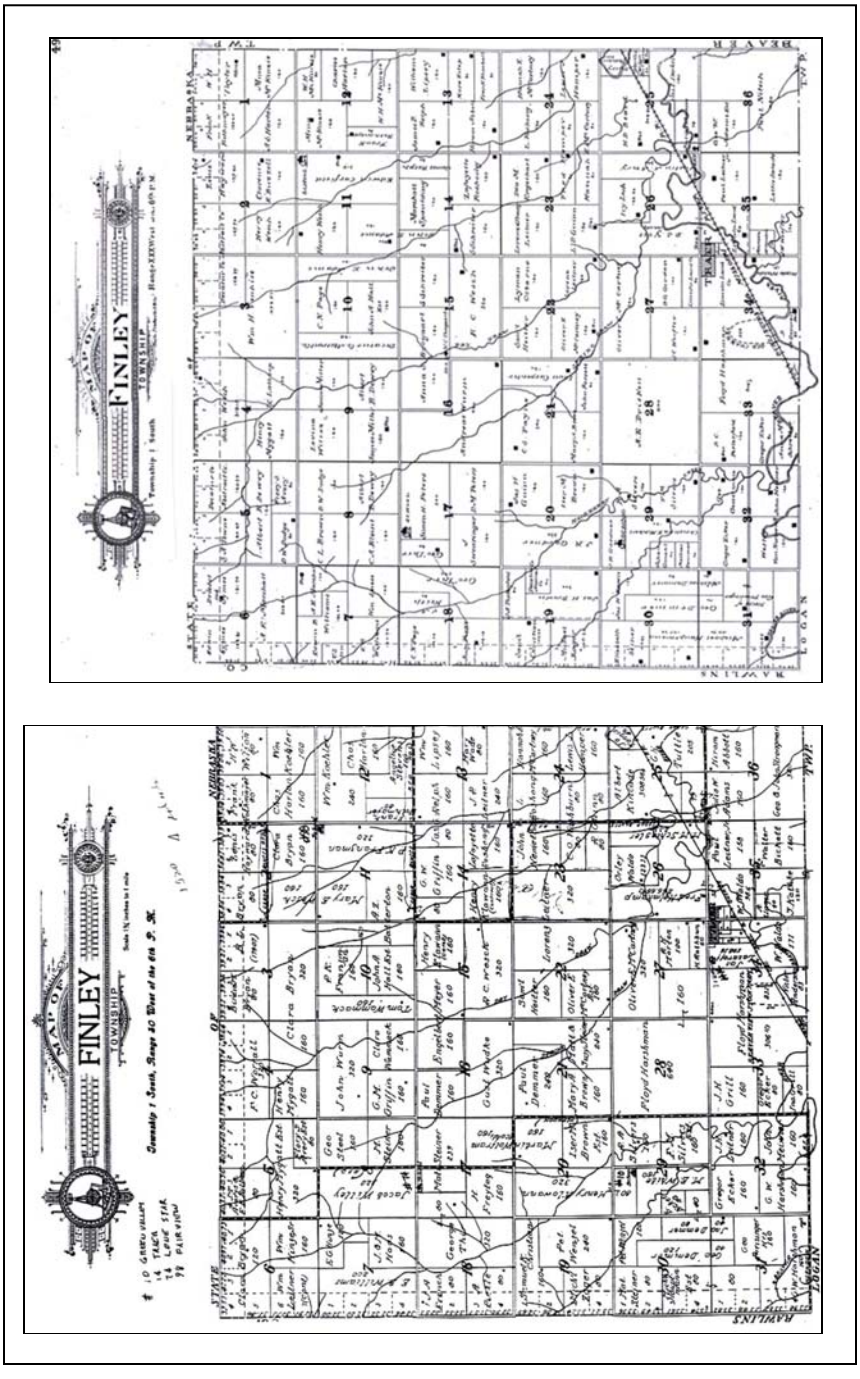

Figure 7. Finley township, Decatur County, 1920 (left) and 1905 (right).

section homesteads a considerable number match exactly and do not need to be altered. For parcels that do not match we manually delete, edit, move, or add lines until farm polygons match the map images, maintaining topology so there are no overlaps or slivers. Finally, we assign unique identifiers and landowner's name to the attribute table of the polygon layer. We have digitized 52 plat maps which will help us place farms and families represented in the census precisely on the landscape.

But ownership obviously reflects a different metric than the operational farm units that are traced by the census. Owners sometimes leased parcels to neighbors or farmers leased parcels owned by non-residents. Thus it was not surprising that four owners appearing on the 1920 plat map could only be located in the census records of adjacent townships. Other spatially specific references showing ownership information include tax assessment records and chains of title for the parcels. These records are available in county court houses across the state, or in rare cases, have been transferred to the state archives and have been used with skill in prior research.[26] While compelling, the ownership phenomena that underlie the land use patterns may have to wait for the next cycle of this research. For now, our desire to place people and farms in the landscapes of our sample townships will rely on plat maps and our interest in examining the physical outcomes of the cumulative trends in household land use histories will focus initially on a series of aerial photographs gathered from the end of our study period. We have acquired air photos for all 25 townships taken in the late 1930s by the Soil Conservation Service from the U.S. National Archives and Record Administration in College Park, Maryland. These images not only provide stunning visual evidence of the scale and scope of the human signature on Kansas grasslands but will inevitably open up further questions and avenues of research.

CONCLUSION

The decision to create a spatially dispersed sample has already begun to inform our view of how agricultural populations developed and adjusted to the grassland. By privileging rural townships with high population densities within a spatially representative sample of environmental contexts, we have confirmed an enduring association between population density and agricultural land use. Not all of these landscapes were initially inviting to newcomers in the 1870s and 1880s. It took nearly two generations for population patterns to mature. But over time population densities did reach certain maxima, bounded by the institutional framework of the land system and by the basic patterns of interaction between population and environment across the state. No matter how much tinkering with boundaries occurred, basic patterns of land use remained strongly associated with population density. As farm families understood what was possible, we expect that the frequent and detailed records of household demography and land use gathered in this study, will question the image of plains agriculture as homogenizing, over-reaching and environmentally improvident.

ACKNOWLEDGEMENTS

The authors wish to thank the many people who contributed to this study. Colleagues at the International Microdata Access Group workshop were generous in sharing their conceptual and programming ideas, especially Peter Tilley. Professor Glenn Deane's statistical advice was invaluable in establishing the parameters of our sample design. Our programmer at ICPSR, Steve Burling, was a model of creativity and good cheer while creating our linking program in Java. Key staff who moved the study moving forward day to day included project manager Shelly Petrinko, research assistant William Myers and GIS analyst Lisa Isgett. Finally, our humblest thanks are extended to a dedicated group of undergraduate research assistants who entered and verified close to a quarter-million records: Andrew Berg, Amber Buycks, Joshua Christopher, Melissa D'Arcy, Erin Garvey, Jared Goldberger, Mike Hanlon, Katie Kiser, Cynthia Lou, Drew Myers, Laura Platt, Margaret Prest and Lily Swanbrow.

ENDNOTES

[1] This research is supported by grant HD44889 from the National Institute of Child Health and Human Development.

[2] N. Ramankutty and J. A. Foley, 'Estimating historical changes in global land cover: Croplands from 1700 to 1992', *Global Biogeochemical Cycles*, 13 (1999), 997–1027 and V. Smil, *Enriching the earth: Fritz Haber, Carl Bosch, and the transformation of world food production* (Cambridge, Mass., 2001).

[3] F. J. Turner, *The frontier in American history* (New York, 1920) and W. P. Webb, *The Great Plains* (New York, 1931).

[4] Great Plains Committee, *The future of the Great Plains: the report of the Great Plains Committee* (Washington, 1936); D. Worster, *Dust Bowl. The Southern Plains in the 1930s* (New York, 1979); R. D. Hurt, *The Dust Bowl: an agricultural and social history* (Chicago, 1981); and J. Opie, *Ogallala: water for a dry land* (Lincoln, Nebraska, 1993).

[5] J. Hightower, *Hard tomatoes, hard times: a report of the agribusiness accountability project on the failure of America's land grant college complex* (Cambridge, Mass., 1973); W. Berry, *The unsettling of America: culture and agriculture* (San Francisco, 1977); L. J. Lear, *Rachel Carson: witness for nature* (New York, 1997); C. Potter, *Against the grain: agri-environmental reform in the United States and European Union* (London, 1998); R. S. Beeman and J. A. Pritchard, *A green and permanent land: ecology and agriculture in the twentieth century* (Lawrence, Kansas, 2001) and S. Stoll, *Larding the lean earth soil and society in nineteenth-century America* (New York, 2002).

[6] M. P. Gutmann and G. Cunfer, *A new look at causes of the Dust Bowl,* Texas Tech University, Publication 99–1 (Lubbock, Texas, 1999).

[7] D. W. Pogue and G. D. Schnell, 'Effects of agriculture on habitat complexity in a prairie-forest ecotone in the southern Great Plains of North America', *Agriculture, Ecosystems and Environment*, 87 (2001), 287–98 and G. Cunfer, *On the Great Plains: agriculture and environment* (College Station, Texas, 2005), 27–9.

[8] J. C. Malin, *The grassland of North America: prolegomena to its history* (Lawrence, Kansas, 1947); J. Diamond, *Guns, germs, and steel the fates of human societies* (New York, 1997); F. Berkes, C. Folke and J. Colding, *Linking social and ecological systems: management practices and social mechanisms for building resilience* (Cambridge, 1998); E. J. Ostrom, E. J. Burger, C. B. Field, R. B. Norgaard and D. Policansky, 'Revisiting the commons: local lessons,

global challenges', Science, 284, (1999), 278–82; E. Lambin, B. L. Turner II, H. J. Geist, S. B. Agbola, A. Angelsen, J. W. Bruce, O. T. Coomes, R. Dirzo, G. Fischer, C. Folke, P. S. George, K. Homewood, J. I. Rik, L. Xiubin Li, E. F. Moran, X. M. Mortimore, P. S. Ramakrishna, J. F. Richards, H. Skanes, G. D. Stone, U. Svedin, T. A. Veldkamp, C. Vogel and J. Xu, 'The causes of land-use and land-cover change: moving beyond the myths', *Global Environmental Change*, 11, (2001), 261–9; B. Donahue, *The great meadow: farmers and the land in colonial Concord* (New Haven, 2004); and J. Diamond, *Collapse: how societies choose to fail or succeed* (New York, 2005).

9 R. L. Forstall, *Population of states and counties of the United States: 1790 to 1990*, U.S. Bureau of the Census (Washington, DC, 1996).

10 T. Snijders and R. Boskers, *Multilevel analysis: an introduction to basic and advanced multilevel modeling* (Thousand Oaks, California, 1999).

11 J. J. Hox and C. J. M. Maas, 'The Accuracy of Multilevel Structural Equation Modeling with Pseudobalanced Groups and Small Samples', *Structural Equation Modeling* 8 (2001), 157–74.

12 J. C. Malin, *Winter wheat in the golden belt of Kansas: a study in adaption to subhumid geographical environment* (Lawrence, Kansas, 1944), preface.

13 E. West, *The contested plains: Indians, goldseekers, and the rush to Colorado* (Lawrence, Kansas, 1998); and W. Brandon, *Quivira: Europeans in the region of the Santa Fe Trail, 1540–1820* (Athens, Ohio, 1990).

14 Kansas, *Statutes*, 1868, ch. 25 § 25.

15 While descriptions of the boundaries of the enumeration districts are available on census microfilm, they seldom furnish enough information to create reliable digital maps or to generate estimates of area.

16 The Public Land Survey System digital files are available free on the web from the Kansas Geospatial Data Commons, <URL: http://www.kansasgis.org/l>. [25 January 2006]

17 Kansas State Board of Agriculture, *Tenth biennial report of the Kansas State Board of Agriculture to the Legislature of the State, for the years 1895 and 1896* (Topeka. 1896).

18 In Doniphan, Pawnee, Pratt, Sedgwick, Smith and Trego counties, the most densely populated rural townships had multiple predecessors or did not represent a smaller version of a larger predecessor.

19 Townships that represented subsets of their larger predecessors included Cottage Grove Township in Allen County, Toledo Township in Chase County, Peoria Township in Franklin County and Voorhees Township in Stevens County (see Table 1).

20 National coverages of gridded historical weather data may be accessed for no charge at PRISM project's website maintained at Oregon State, <URL: http://www.ocs.orst.edu/prism/>. [25 January 2006].

21 County level coverages of digital soil maps may be downloaded at no charge from the United States Department of Agriculture's Natural Resources Conservation Service website, through its Soil Data Mart portal, <URL: http://soildatamart.nrcs.usda.gov/>. [25 January 2006].

22 J. R. Anderson, E. E. Hardy, J. T. Roach and R. E. Witmer, *A land use and land cover classification system for use with remote sensor data*. Geological Survey Professional Paper 964 (Washington, DC, 1976).

23 The Minnesota Population Center generously shared a data entry program written by their programmers to enter a public use microdata sample of the 1910 federal census. Steve Burling, a programmer at ICPSR, modified MPC's program to our needs.

24 G. J. Guth, 'Surname Spellings and Computerized Record Linkage', *Historical Methods Newsletter*, 10 (1976), 10–19; G. Bouchard, P. Bérard and Y. Lavoie, 'FONEN: un code de transcription phonétique pur la reconsititution automatique des familles sanguenayennes', *Population*, 36 (1986), 1085–1103; D. De Brou and M. Olsen, 'The Guth algorithm and the nominal record linkage of multi-ethnic populations', *Historical Methods*, 19 (1986): 20–4; M. Fair, E. P. Lalonde and H. B. Newcombe, 'Application of exact ODDS for partial agreements of names in record linkage', *Computers and Biomedical Research*, 24 (1991), 58–71; J. E. Vetter, J. R. Gonzalez and M. P. Gutmann, 'Computer-assisted record linkage using a relational database system', *History and Computing*, 4 (1992), 34–51; I. Winchester, 'What

every historian needs to know about record linkage for the microcomputer era', *Historical Methods*, 25 (1992), 149–65; P. Adman, S. W. Baskerville and K. F. Beedham, 'Computer-assisted record linkage: or how best to optimize links without generating errors', *History and Computing* 4 (1992), 2–15; C. Harvey, E. M. Green and P. J. Corfield, 'Record linkage theory and practice: an experiment in the application of multiple pass linkage algorithms', *History and Computing*, 8 (1994), 78–89; P. Tilley and C. French, 'Record linkage for nineteenth-century census returns: automatic or computer aided?', *History and Computing*, 9 (1997), 122–33; S. I. Hautaniemi, D. L. Anderton and A. C. Swedlund, 'Methods and validity of record linkage: matching 1850–1912 census and death records in two Massachusetts towns', *Historical Methods*, 33 (2000), 16–29; and P. Tilley, 'Creating life histories and family trees from nineteenth-century census records, parish registers and other sources, *Local Population Studies*, 68 (2002), 63–81.

[25] Malin, *Grassland of North America*, 1947; D. A. McQuillan, *Prevailing over time: ethnic adjustment on the Kansas prairies, 1875–1925* (Lincoln, Nebraska, 1990); G. Bouchard, 'Family reproduction in new rural areas: outline of a North American model', *Canadian Historical Review*, 75 (1994), 475–510; G. Bouchard, *Quelques arpents d'Amérique: population, économie, famille au Saguenay, 1838–1971* (Montréal, 1996); J. Gjerde, *The minds of the West: ethnocultural evolution in the rural Middle West, 1830–1917* (Chapel Hill, North Carolina, 1997) and K. M. Sylvester, *The Limits of rural capitalism: family, culture and markets in Montcalm, Manitoba, 1870–1940* (Toronto, 2001).

[26] McQuillan, *Prevailing over time* and R. K. Loewen, *Family, church, and market: a Mennonite community in the Old and New worlds, 1850–1930* (Urbana, Illinois, 1993).

HISTORY and COMPUTING

Volume 14
(Nos 1 and 2 2002)
PUBLISHED 2006

EDITOR
Matthew Woollard

GUEST EDITORS
Lisa Y. Dillon and Evan Roberts

Published by Edinburgh University Press
for the
Association for History and Computing

CONTENTS

CONTRIBUTORS

Anders Brändström is Professor of Historical Demography at Umeå University, and Director of the Demographic Database. His research interests include the infant mortality decline in Sweden and the Nordic countries, demography of nineteenth century hospital patients, social mobility in nineteenth century Sweden, and marriage patterns in eighteenth and nineteenth century Sweden.

Cameron Campbell is Professor of Sociology at the University of California at Los Angeles. With James Lee, he is the author of *Fate and fortune in rural China: social organization and population behavior in Liaoning, 1774–1873*. He is one of the co-authors of *Life under pressure: mortality and living standards in Europe and Asia, 1700–1900*.

Geoff Cunfer is Associate Professor of History at the University of Saskatchewan where he specializes in environmental history, the American Great Plains, and Geographic Information Systems (GIS) methods for history. He taught for six years at Southwest Minnesota State University and is the author of *On the Great Plains: agriculture and environment* (2005). 'Scaling the dust bowl' is forthcoming in Anne Kelly Knowles, ed., *Placing history: how maps, spatial data, and GIS are changing historical scholarship*.

Ros Davies is Senior Research Associates at the Cambridge Group for the History of Population and Social Structure, Geography Department, Cambridge University.

Lisa Dillon is an assistant professor at the Département de Démographie, Université de Montréal. She supervises several projects to create, refine and

History and Computing, 14 (1+2) 2002 (publ. 2006), v–viii ISSN 0 957-0144

use nineteenth-century Canadian census microdata, and teaches courses in the history of trans-Atlantic populations, historical demographic methods and population politics. Dillon's research focuses on intergenerational relations, concepts of age and life course transitions in eighteenth- and nineteenth-century Québec, Canada and the United States.

Gordon Darroch is Professor of Sociology and Senior Scholar, York University, Toronto. He is a co-leader of the Canadian Century Research Infrastructure Project, a national, collaborative project constructing an integrated public-use database from the Canadian historical censuses. He has published widely on substantive and methodological issues in historical microdata.

Sören Edvinsson is an Associate Professor of History at Umeå University, where he specializes in research in Swedish demographic history and the history of public health. He has published widely on these topics, including numerous papers on mortality and life expectancy, and methods for the study of demographic history.

Eilidh Garrett is Senior Research Associates at the Cambridge Group for the History of Population and Social Structure, Geography Department, Cambridge University.

Danielle Gauvreau is a social demographer specializing in the history of the Quebec population. She teaches in the Department of Sociology and Anthropology at Concordia University and has a forthcoming book on the Quebec fertility transition (with Diane Gervais and Peter Gossage). She is currently working on a SSHRC funded project entitled 'Adaptations of youth in the industrializing city: Montreal, 1840–1900' (with Sherry Olson).

Myron P. Gutmann is Professor of History and Director of the Inter-university Consortium for Political and Social Research at the University of Michigan. He has broad interests in interdisciplinary history, especially population, health, immigration, and environment, in both Europe and the United States. He has published *War and rural life in the early modern Low Countries* (1980), *Towards the modern economy* (1988), and more than sixty articles and chapters.

James Lee is Professor of History and Sociology at the University of Michigan, and Director of the Center for Chinese Studies. With Wang Feng, he co-authored *One-quarter of humanity: Malthusian mythology and Chinese reality*. Most recently, he was one of the co-authors of *Life under pressure: mortality and living standards in Europe and Asia, 1700–1900*.

Susan Hautaniemi Leonard is Assistant Research Scientist at the Inter-university Consortium for Political and Social Research at the University of Michigan. An anthropologist with broad interests in population and environment research including historical epidemiology, mortality, aging, and household dynamics in the United States, she has published in *Social Science History*, *Journal of the History of Medicine and Allied Sciences*, *Historical Methods*, *Human Biology* and *American Journal of Human Biology*.

Kees Mandemakers works at the International Institute for Social History (IISH) as senior research fellow, heading the Historical Sample of the Netherlands (HSN). Main research interests: methodology of large historical databases (topics: data processing, sample designs), social stratification and mobility, social history (topics: secondary education, infant and child mortality).

Alice Reid is Senior Research Associates at the Cambridge Group for the History of Population and Social Structure, Geography Department, Cambridge University.

Evan Roberts is a PhD candidate in History at the University of Minnesota. He co-ordinates the North Atlantic Population Project, which harmonizes historical censuses from North America and Europe. His research focuses on the economic history of the American labour market, and comparative social history.

Steven Ruggles is Distinguished McKnight University Professor of History, and Director of the Minnesota Population Center at the University of Minnesota. He has been instrumental in the development of public samples of census microdata. His research focuses on living arrangements of the elderly and household composition in the United States.

Kenneth M. Sylvester is Assistant Research Scientist at the Inter-university Consortium for Political and Social Research at the University of Michigan. An historian with interdisciplinary interests in agriculture, environment, and population research, in Canada and the United States, he has published *The limits of rural capitalism* (2001) and is co-author of the forthcoming 'Dustbowl legacies: long term change and resilience in the shortgrass steppe' in Charles Redman and David Foster, eds., *Agrarian landscapes in transition*.

Patricia Thornton is a population and historical geographer at Concordia University, Montreal. She has written extensively (with Sherry Olson) on nineteenth century Montreal (infant mortality, cultural differences in demographic behaviour and the Irish). She has also written on nineteenth century out-migration from Canada to the United States. She is currently working on a

SSHRC funded project entitled 'Adaptations of youth in the industrializing city: Montreal, 1840–1900' (with Sherry Olson).

Pär Vikström is the Director of Information Technology at the Demographic Database at Umeå University. He has twenty-three years experience in working with longitudinal databases, and has been involved in many research projects with Umeå University researchers.

ABSTRACTS

Introduction: Longitudinal and cross-sectional historical data: intersections and opportunities
Lisa Dillon and Evan Roberts

Historians and other social scientists have become increasingly interested in the construction of historical, longitudinal datasets in the past decade. This special issue of *History and Computing* brings together eight articles from a recent workshop on methods, materials and results from historical, longitudinal data. We summarize the contribution of these articles, and connect their themes and results. In particular, we discuss the trade-offs between accuracy, representativeness and sample size in creating longitudinal databases from cross-sectional sources. Procedures for linking individuals across different data sources vary among the different projects in this issue, with implications for the analyses that can be performed. Another set of papers in this issue discuss the construction of longitudinal databases from explicitly longitudinal sources, such as population and event registries. Although these sources are longitudinal in their original design, like the cross-sectional data sources, differences in the social context of data collection affect the type of data that was collected. Despite these intrinsic research challenges, the analysis of historical, longitudinal databases have contributed important new findings in demographic and economic history, and will continue to do so.

State views and local views of population: linking and comparing genealogies and household registers in Liaoning, 1749–1909
Cameron Campbell and James Lee

To assess the consistency of recording in historical Chinese data, we compare the demography of the same families recorded in two separate sources: household

History and Computing 14 (1+2) 2002 (publ. 2006), 225–230 ISSN 0 957-0144

registers and lineage genealogies. We find that these sources offer contradictory accounts. Reflecting differences between official and private constructions of kinship, individuals recorded in the household registers are often missing from the genealogies. We apply multivariate techniques to examine how the families that compiled genealogies differed from other families, and how the individuals recorded in genealogies differed from unrecorded members of the same family. We find that the families for which genealogies are available are highly selective in terms of their socioeconomic standing, and that there is additional selectivity in terms of which members were recorded. We also find that fertility estimates from genealogies may be biased upward, rather than downward, because genealogies are more likely to omit adults who have few or no offspring. We conclude by discussing the implications for interpretation of published findings from sources such as lineage genealogies and household registers that have been the mainstay of historical studies of Chinese kinship and demography.

Demography and environment in grassland settlement: using linked longitudinal and cross-sectional data to explore household and agricultural systems

Kenneth M. Sylvester, Susan Hautaniemi Leonard, Myron P. Gutmann and Geoff Cunfer

The Demography and Environment in Grassland Settlement project (DEGS) is a study of the relationship between population and environment in Kansas during its settlement and conversion from grassland to grain cultivation and rangeland. The research team involved in this project had as its goal to bring together data about farms and farm families in order to understand the core transformations in land use and family dynamics that took place during the process of settling and developing an agricultural landscape. Kansas is ideally suited for this study by virtue of its location, history, and the documents that exist about it. We are assembling a linked database of farm and family census records for twenty-five townships scattered across the state. This paper is about the process of choosing that sample, about the data we have accumulated, and about the process we are undertaking to link records about families and farms through time, and to attempt to find their locations in space. The paper also reports on the creation of a semi-automated linkage program in Java. The program allows users to view records in tabular form, one close to the original historical documents, and compare information across the agricultural and population censuses, and make decisions confirming or ignoring the suggested links displayed when employing the matching algorithms in the program. A table preserving linked persons and households, in the same year or over time, is then is stored separately in a relational database with a set of unique time-person identifiers.

Nineteenth-century Scottish demography from linked censuses and civil registers: a 'sets of related individuals' approach
Alice Reid, Ros Davies and Eilidh Garrett

This paper reports on a project to link nineteenth century Scottish civil registration and census records, broadening and extending linking techniques with a new emphasis on using sets of individuals related through families, households, and neighbourhoods to establish links. The project aims to do for Scotland what cannot yet be done for England: to link the comprehensive civil register records of births, deaths and marriages in an area with individuals in the census returns covering the same area to form a reconstitution-like data set. The resulting data set will be richer than reconstitution however, due to the additional information in the Scottish sources (for example, maiden name is routinely given for mothers in civil registers, and birth records also provide parents' place of marriage); censuses yield information relating to household size and structure and also provide decennial observations of those individuals who do not experience any demographic event in the inter-censal period. The paper illustrates the linkage technique in the case of the Island of Skye, and speculates on the prospects for it's successful application to urban communities such as the lowland town of Kilmarnock.

Building life course datasets from population registers by the Historical Sample of the Netherlands (HSN)
Kees Mandemakers

The Historical Sample of the Netherlands (HSN) compiles life course data for the nineteenth and twentieth century population. Most of the data for the construction of life courses are extracted from the population registers that allow us to trace persons from the cradle to the grave. In this article, we will present the way how the HSN reconstructs raw material from the Dutch population registers into output files that can be used by researchers. This process is comprised of several phases, of which the most important is linking all different appearances of the same persons across all entries, dating all individual arrivals and departures, reconstituting households and adding individuals to these households, dating all events related to individuals and households, and finally outputting the data in a new data structure, at the same time removing all redundant information. Discussion of the HSN software is situated in the context of the character and main sources of the HSN database, the population registers, and the variables that can be distilled out of them.

Longitudinal databases – sources for analyzing the life-course: characteristics, difficulties and possibilities
Pär Vikström, Sören Edvinsson and Anders Brändström

The Demographic Database at Umeå University has over the past thirty years constructed large longitudinal databases of the Swedish population from the eighteenth century, based on material in parish registers. The Demographic Database has developed into one of the world's most significant historical, longitudinal data sources. In this paper we discuss some of the advantages and problems of constructing longitudinal databases, and applying them in research. One of the principal problems in longitudinal data based on registers is the inconsistency of information available about individuals, either over time, or between different sources. A solution to this problem that the Demographic Database applies has been the conceptual distinction between source variables, and analysis variables derived from the source according to well-defined rules.

Reconciling cross-sectional and longitudinal measures of fertility, Quebec, 1890–1900
Patricia A. Thornton and Danielle Gauvreau

In the absence of vital registration, studies of the onset and early phases of the fertility transition in North America have been seriously hampered and yet the seemingly early timing of the decline, the multi-ethnic nature of the population and continuous flow of immigrants from Europe suggest that North America has much to offer to this debate. This paper is primarily methodological. By reconciling cross-sectional census measures of fertility using the own-child method (1901) with those derived from a longitudinal ten-year panel (1891–1901) using family reconstitution, for Montreal and surrounding region it exposes some of the weaknesses and the potentials of the two methods most often currently used and the advantages of combining methods. Own-children measures of marital fertility are seriously affected by significant local differences in infant survival between rural and urban areas and between cultural groups as well as by residual effects of duration and timing of marriage, while small-scale longitudinal studies in complex environments cannot always render reliable results for all sub-populations nor can they necessarily be 'scaled up'. They suggest that national and even regional averages of fertility may conceal large diversity, which in turn raises questions about the existence of any single transition with uniform characteristics and timing, or universal cause. Instead we argue different groups in different environments may actually have been fine-tuning their fertility behaviour to compensate for the differential effects of mortality through adjustments to both marriage and fertility within marriage.

Semi-automated record linkage with surname samples: a regional study of 'case law' linkage, Ontario, 1861–1871
Gordon Darroch

With rare exception, North America studies linking nominative historical records beyond single communities have relied on the U.S. census indexes. This paper reports an alternative, semi-automated approach to linkage developed for a large region of South-Central Ontario for censuses of 1861 and 1871. In principle, the procedure applies to any records groups. It combines unique surname samples with linkage procedures based computer sorting and manual matching of a limited number of largely time-invariant variables, but risks selection bias by further enlisting the records of other members of the dwellings of those traced. Labelled "case-law" linkage, final linkage judgements are based on specified rules supplemented by generalized case examples of their many possible permutations in practice. The approach trades off absolute consistency in linkage against saving the considerable costs of developing a fully automated solution for large, complex files. We examine implications of surname samples and selection bias in linkage. Comparisons of linked and baseline (1861) samples suggests quite predictable, but generally modest, differences. The linkage procedure tends to under-represent solitaries and the tiniest households, but multivariate analyses, using dwelling size in 1861 as a proxy for multiple records in the linkage, suggests that this selection bias is modest.

Challenges and opportunities for census record linkage in the French and English Canadian context
Lisa Dillon

This article addresses a new initiative at the Département de Démographie, Université de Montréal, to develop a historical demography research infrastructure which combines cross-sectional and longitudinal population microdata in an effort to better address demographic and family change during the nineteenth century. Here we discuss the record linkage procedures used in a project to link married couples and single persons from the 1 per cent sample of the 1871 Canadian census to the 100 per cent database of the 1881 Canadian census. Following an analysis of differences in the Canadian and U.S. historical populations, differences with implications for record linkage, the article discusses the treatment of Anglophone and Francophone names and describes the linkage criteria used to link persons from 1871 to 1881, beginning with married couples. Linked cases tended to be younger, were more likely to be non-migrants and were more likely to be farmers than the general population of married persons in 1871. To avoid some of the particular biases introduced by the automatic record linkage procedure, it will be necessary to use the automatically-linked cases in conjunction with the manually-linked cases.

Linking historical censuses: a new approach

Steven Ruggles

This paper describes the Minnesota Population Center's ongoing project to create samples of individuals and families linked from samples of the 1860–1870 and 1900–1930 censuses to a complete-count dataset of the 1880 American census. Because of the large populations to be linked, the MPC samples will prioritize creating representative samples rather than maximising the number of linked individuals. This approach differs from several other linking projects, and offers some advantages for researchers.