Journal of Data Science **10**(2012), 579-595

A Model for Spatially Disaggregated Trends and Forecasts of Diabetes Prevalence

Peter Congdon
Queen Mary University of London

Abstract: A multilevel model (allowing for individual risk factors and geographic context) is developed for jointly modelling cross-sectional differences in diabetes prevalence and trends in prevalence, and then adapted to provide geographically disaggregated diabetes prevalence forecasts. This involves a weighted binomial regression applied to US data from the Behavioral Risk Factor Surveillance System (BRFSS) survey, specifically totals of diagnosed diabetes cases, and populations at risk. Both cases and populations are disaggregated according to survey year (2000 to 2010), individual risk factors (e.g., age, education), and contextual risk factors, namely US census division and the poverty level of the county of residence. The model includes a linear growth path in decadal time units, and forecasts are obtained by extending the growth path to future years. The trend component of the model controls for interacting influences (individual and contextual) on changing prevalence. Prevalence growth is found to be highest among younger adults, among males, and among those with high school education. There are also regional shifts, with a widening of the US "diabetes belt".

Key words: Context, diabetes, forecasts, prevalence, risk factor.

1. Introduction

A number of nationwide forecasts of diabetes prevalence in the US have been produced, and predict a continued rise in prevalence, related to factors such as rising obesity and differential growth in minority groups more prone to the condition (Huang *et al.*, 2009; Boyle *et al.*, 2010). However, there are wide variations in diabetes prevalence between different parts of the US (e.g., Barker *et al.*, 2011; Ford *et al.*, 2005), and geographically disaggregated forecasts are important for planning public health interventions.

This paper describes a method for analyzing recent geographic trends in prevalence using health survey data, and for projecting those trends into the future. The model used includes parameters to represent the impact on cross-sectional

prevalence variation of both socio-demographic categories and geographic context of place of residence. Once enduring influences on prevalence are controlled for, the remaining component of the model acts to estimate contextual and person level influences on prevalence growth rates.

While major influences on diabetes prevalence levels are well established, such as race and age gradients, evidence on significant influences on change in prevalence is less clear. After allowing for the overlapping effects on changing prevalence of factors such as age, gender, race, education, and geographic context, there is little evidence on which of these factors are the significant drivers of the growth in prevalence.

The main source data used relates to trends in diagnosed diabetes from the Behavioral Risk Factor Surveillance System (BRFSS) survey, an annual survey of chronic disease prevalence in the US that has included regular questions on diagnosed diabetes. This is supplemented by evidence from the National Health and Nutrition Examination Survey (NHANES) on the ratio of undiagnosed to diagnosed diabetes (Cowie *et al.*, 2006), so that forecasts of the total diabetes burden can be obtained.

2. Risk Categories and Model Components

The analysis here is based on observed prevalence of diagnosed diabetes from 2000 through to 2010, as provided by eleven BRFSS surveys. Prevalence differences and growth are modelled according to categories defined by socio-demographic risk (or protective) factors of individuals, and also according to the geographic context of their place of residence. The model can be extrapolated into the short term future to provide geographically disaggregated forecasts, here for 3141 US counties. To facilitate aggregation of county estimates and forecasts, prevalence rates are expressed as a product of a contextual (geographic) relative risk and a rate reflecting individual risk factors only.

The model is based on data relating to totals of diabetes cases (y), and total populations, (n). These totals are disaggregated according to calendar year, individual risk/protective attributes, and residence category. Individual risk factors are gender, age (age groups 18-49, 50-64 and 65+), race (white non-Hispanic, black non-Hispanic, Hispanic, or other) and education categories (less than high school, high school graduate, some college, or college graduate).

Contextual factors are US census division (see http://www.eia.gov/emeu/reps/maps/us_census.html), and the poverty level of the county of residence, namely the poverty quartile of the county of residence, as defined by annual small area income and poverty (SAIPE) estimates from the US Census Bureau. County of residence is not provided for all survey subjects: for example, in the 2010 BRFSS survey, 90.5% of records had identified counties. For developing the

county poverty quartile, a preliminary regression is therefore undertaken to estimate county poverty when the county is not identified in the survey. This regression is based on those survey records where county of residence is identified. A linear regression of the log county poverty rate is made on state poverty and on individual BRFSS respondent level race and education category. This regression is used to estimate county poverty (and hence assign poverty quartile) for those records where county of residence is not identified.

The main longitudinal prevalence model (see Section 3) then has two components: (a) cross-sectional parameters to represent the impact on prevalence of socio-demographic categories (e.g., age, race) and geographic context, namely to explain persisting inequalities in prevalence common to all years; and (b) trend parameters to represent the impact on changing prevalence (effectively prevalence growth) of socio-demographic categories and geographic context.

Cross-sectional variations in prevalence according to the above risk factors and geographic categories are well established. A gradient in diabetes prevalence by age is reported by Mokdad *et al.* (2001), while Maty *et al.* (2005) and Smith (2007) find that socioeconomic disadvantage, especially low educational attainment, is a significant predictor of incident Type 2 diabetes. As to race differentials, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK, 2011) report diagnosed prevalence rates among adults (in 2007-09) of 7.1% for white non-Hispanics, 12.6% for black non-Hispanics, and 11.8% for Hispanics. The wide regional contrasts in diabetes prevalence are reviewed by Barker *et al.* (2011) who identify a "diabetes belt" in the US south east. Hence the effects of the above risk or residence categories on prevalence levels are all likely to be significant in terms of a regression predicting enduring prevalence contrasts.

However, socio-demographic and contextual factors are not necessarily all significant influences on changes in the pattern of diabetes prevalence. There is relatively little evidence on whether inequalities in diabetes prevalence are growing or diminishing primarily according to age, or race, or education, or geographic context after allowing for the overlapping effects on changing prevalence of these factors. For example, changing regional prevalence differences may actually reflect impacts on changing prevalence of regional differences in race composition, proportions with college education or county poverty levels.

3. Regression Methods

To model influences on prevalence variations and growth, a binomial regression is used with prevalence probabilities specific to time, socio-demographic status, and geographic context. The goal of the model is in part description of existing levels and trends, but additionally to provide a scheme for geographically disaggregated forecasts. Let the model categories be denoted g $(= 1, 2$ for

males, females), d (= 1 to 9) for divisions, q $(= 1, \cdots, 4)$ for county poverty quartile, a $(= 1, \cdots, 3)$ for age band, r $(= 1, \cdots, 4)$ for race, e $(= 1, \cdots, 4)$ for education level, and t $(= 1, \cdots, 11$ for years 2000 to 2010). The BRFSS provides totals $y_{gdqaret}$ of diagnosed diabetes cases, and totals $n_{gdqaret}$ of adult survey subjects (aged 18 and over) in the relevant risk categories. There are a maximum of 38016 $(= 2 \times 9 \times 4 \times 3 \times 4 \times 4 \times 11)$ observations, based on all possible combinations of categories. A weighted likelihood is used with total survey weights $w_{gdqaret}$ aggregated over relevant survey subjects, so that the weighted log-likelihood for a particular risk category combination is

$$w_{gdqaret} \times \{\log(n_{gdqaret}!) - \log(y_{gdqaret}!) - \log([n_{gdqaret} - y_{gdqaret}]!) \\ + y_{gdqaret} \log(\pi_{gdqaret}) + (n_{gdqaret} - y_{gdqaret}) \log(1 - \pi_{gdqaret})\},$$

where $\pi_{gdqaret}$ is the probability of diagnosed diabetes specific to time, socio-demographic category, and geographic context.

A binomial regression is applied using the Bayesian package WINBUGS and Markov Chain Monte Carlo (MCMC) sampling. The regression uses a log link (Blizzard and Hosmer, 2006) to predict $\pi_{gdqaret}$. When assessing the effect of a particular predictor, it is of interest to estimate the relative risk for that predictor adjusted for the effects of the other predictors. When prevalence of an event is low, the odds ratio provides a good approximation to the relative risk (Agresti, 2002), and the usual logit-link binomial model can be used to estimate relative risks. If the event probability is not small (as in diabetes at older ages), then a log-link binomial model can be used to directly estimate the relative risk. Using the log link does not ensure that predicted probabilities are mapped to the [0,1] range, and so a constraint is used to ensure prevalence probabilities $\pi_{gdqaret}$ do not exceed 1, though this is only an issue very early in the MCMC sampling. In the particular application here, use of a log link ensures that one can express each time-specific prevalence rate $\pi_{gdqaret}$ as the product of a contextual relative risk (the impact of geographic context) and a rate reflecting individual risk factors only (see Section 4). This separation is not possible straightforwardly with a logit link.

The model for the prevalence rates includes (a) parameters for enduring cross-sectional differences, and (b) parameters that explain varying prevalence growth rates between different socio-demographic groups and geographic contexts. Let cross sectional parameters be denoted α_1 (for intercept), α_{2g} (for gender effects), α_{3d} (for division effects), α_{4q} (for the effects of county poverty quartile), α_{5a} (for age effects), α_{6r} (for race effects), and α_{7e} (for effects of education level).

A linear growth trend (in log prevalence) is assumed, which facilitates out-of-sample forecasting, and seems reasonable in view of actual prevalence trends between 2000 and 2010. For example, Figure 1 depicts changes in logarithmically

transformed adult prevalence (log of age standardised percentages) for the US and the nine census divisions between 2000 and 2010. Then let parameters affecting change in prevalence be: δ_1 (for average growth in log-linear scale), δ_{2g} (for differential prevalence growth by gender), δ_{3d} (for differential growth in prevalence by division), δ_{4q} (for effects on prevalence growth of county poverty quartile), δ_{5a} (for age effects on growth), δ_{6r} (for race effects on growth), and δ_{7e} (for effects on growth of education level).

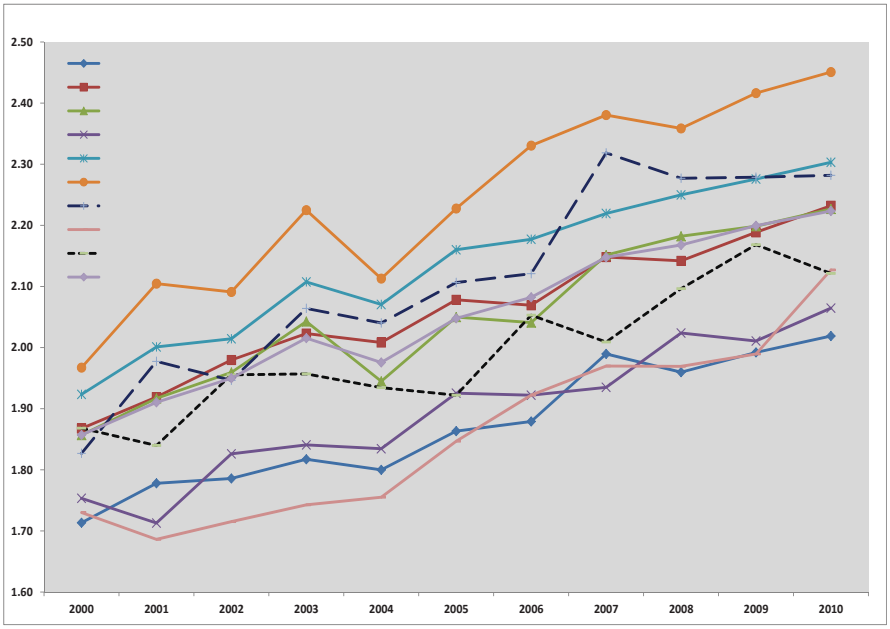

Figure 1: Trend in adult diabetes prevalence (log percent diagnosed rate)

Then the regression model has the form

$$\log(\pi_{gdqaret}) = \alpha_1 + \alpha_{2g} + \alpha_{3d} + \alpha_{4q} + \alpha_{5a} + \alpha_{6r} + \alpha_{7e} + (\delta_1 + \delta_{2g} + \delta_{3d} + \delta_{4q} + \delta_{5a} + \delta_{6r} + \delta_{7e})t^*,$$

where $t^* = (t-1)/10$. The time unit in the trend (and forecast) component is then decadal units t^*, with $t^* = 0$ for 2000, $t^* = 1$ for 2010, $t^* = 1.1$ for 2011 and $t^* = 1.5$ for 2015.

All except the division parameters $\{\alpha_{3d}, \delta_{3d}\}$ are treated as fixed effects, and constrained to sum to zero to ensure identification (rather than corner constraints). For example, whereas a corner constraint sets $\alpha_{61} = 0$, here α_{61} is obtained as $\alpha_{61} = -[\sum_{r=2}^{4} \alpha_{6r}]$, with parameters $\{\alpha_{62}, \cdots, \alpha_{64}\}$ as unknowns. All unknown fixed effects parameters are assigned $N(0, 100)$ priors. The division parameters $\{\alpha_{3d}, \delta_{3d}\}$ are taken as random effects, and follow a conditional

autoregressive structure (Besag *et al.*, 1991) that reflects their likely spatial correlation, as illustrated by evidence regarding the "diabetes belt" in the US. For example, the α_{3d} terms condition on $\alpha_{3[d]}$ in all other divisions, according to

$$\alpha_{3d}|\alpha_{3[d]} \backsim N(\bar{A}_d, \frac{\phi_\alpha}{L_d}),$$

where $\bar{A}_d$ is the average of the α_{3c} in the $c = 1, \cdots, L_d$ divisions contiguous to the d^{th} division, and ϕ_α is a variance parameter. To ensure identifiability the α_{3d} and δ_{3d} are centred at each iteration. Gamma priors with shape 1 and scale 0.01 are assumed for the inverse variances of the (α_{3d}) and (δ_{3d}).

Model checks are applied using predictive replicates, generically y_{rep} (here specific for time, socio-demographic category, and geographic context) with posterior predictive distribution

$$p(y_{rep}|y) = \int p(y_{rep}|\theta)p(\theta|y)d\theta,$$

where θ collectively denotes model parameters. For fraction α (e.g., $\alpha = 0.1$), one would expect approximately $(1-\alpha)$ of the observations $y_{gdqaret}$ to fall within the $(1-\alpha)$ predictive interval, namely the $(1-\alpha)\%$ credible interval for $y_{rep,\, gdqaret}$ (Gelfand, 1996, p. 153).

4. Obtaining Spatially Disaggregated Prevalence Estimates and Forecasts

By virtue of the form of the model and the log link, one can express each time-specific prevalence rate $\pi_{gdqaret}$ as the product of a contextual relative risk and a rate reflecting individual risk factors only. Here the context is provided by nine US census divisions and US counties (there are 3141 counties across the US). The contextual relative risk combines the impact of division and county poverty quartile parameters, namely

$$\rho^{ctx}_{d,q,t} = \exp(\alpha_{3d} + \alpha_{4q} + \delta_{3d}t^* + \delta_{4q}t^*),$$

while the US-wide rate for the impact of individual risk factors (sex-age-race-education) is

$$\pi^{ind}_{garet} = \exp(\alpha_1 + \alpha_{2g} + \alpha_{5a} + \alpha_{6r} + \alpha_{7e} + [\delta_1 + \delta_{2g} + \delta_{5a} + \delta_{6r} + \delta_{7e}]t^*).$$

The contextual effect $\rho^{ctx}_{d,q,t}$ allows county prevalence estimates to be made that reflect prevalence variations (relative to the national sex-age-race-education benchmark π^{ind}_{garet}) according to the division the county is located in, and according to county poverty level. Thus if county c is located in division d_c and

in county poverty quartile q_c, its sex-age-race-education prevalence schedule is estimated as

$$\pi_{cgaret} = \rho^{ctx}_{t,d_c,q_c}\pi^{ind}_{garet}.$$

Forecasts beyond the last observed survey year 2010 to 2011 and years thereafter can be made by setting $t = 11, 12, \cdots$ (i.e. $t^* = 1.1, 1.2, \cdots$) in the model set out in Section 3,

$$\begin{aligned}\log(\pi_{gdqaret}) = \alpha_1 + \alpha_{2g} + \alpha_{3d} + \alpha_{4q} + \alpha_{5a} + \alpha_{6r} + \alpha_{7e} \\ + (\delta_1 + \delta_{2g} + \delta_{3d} + \delta_{4q} + \delta_{5a} + \delta_{6r} + \delta_{7e})t^*.\end{aligned}$$

The above decomposition to obtain π_{cgaret} can then be applied for these future years.

However, the interest is typically in summary county level estimates (and forecasts), or estimates for demographic variables such as age, race and sex. To average over education categories, and hence obtain estimated age-sex-race or age-sex prevalence rates by county, let $\{w^{educ}_{ce}, e = 1, \cdots, 4\}$ with $\sum_{e=1}^{4} w^{educ}_{ce} = 1$ ($e = 1$ for elementary schooling only through to $e = 4$ for college graduate) denote the county's education mix. Let

$$\rho^{educ}_{c,t} = \exp\left[\sum_{e=1}^{4}(\alpha_{7e} + \delta_{7e}t^*)w^{educ}_{ce}\right],$$

be relative prevalence risks summarising the effect on prevalence of a county's education composition. Ideally year-specific county education composition data w^{educ}_{cet} would be available, but intercensal estimates of county education mix are not available, and here 2000 Census data are used. Also let

$$\pi^{ind}_{gart} = \exp(\alpha_1 + \alpha_{2g} + \alpha_{5a} + \alpha_{6r} + [\delta_1 + \delta_{2g} + \delta_{5a} + \delta_{6r}]t^*)$$

be US-wide prevalence schedules for sex-age-race category. Then county sex-age-race prevalence estimates that reflect both the county's education mix, and its division and poverty level, are obtained as

$$\pi_{cgart} = \rho^{ctx}_{d_c,q_c,t}\rho^{educ}_{c,t}\pi^{ind}_{gart}.$$

Suppose further, one seeks simple age-sex prevalence rates for counties, aggregating over race. For example, since diabetes prevalence is higher for black non-Hispanics, counties with above average proportions of black people would be expected to have higher prevalence by virtue of their racial composition. Let $\{w^{race}_{crt}, r = 1, \cdots, 4\}$ with $\sum_{r=1}^{4} w^{race}_{crt} = 1$ ($r = 1$ for white non-Hispanics through

to $r = 4$ for other ethnic groups) denote the county's race mix in year t. Then weighted averages

$$\pi_{cgat} = \sum_{r=1}^{4} \pi_{cgart} w_{crt}^{race},$$

provide county level age-sex prevalence estimates that take account of a county's education and ethnic composition, and its geographic context (division, poverty quartile). Here race mix is available from county population estimates up to 2010, but is assumed constant thereafter as population forecasts for US counties are not currently made.

5. Results

Inferences on the parameters $\{\alpha, \delta\}$ are based on the second halves of two chain runs of 10,000 iterations with convergence by iteration 5000 assessed using Gelman-Rubin statistics. Model checks are satisfactory: 92.4% of the observations are within 90% credible intervals of $y_{rep,\, gdqaret}$.

Table 1 shows estimates for the cross-sectional α-parameters, and how they translate into relative risks that enhance or diminish the overall prevalence rate. The cross-sectional parameters notionally represent the start-point of the time range in the observed data, namely 2000. Rates for demographic categories are obtained by exponentiating: for example, the rate for White non-Hispanics (WNH) aged 18-49 is $\exp(\alpha_1 + \alpha_{51} + \alpha_{61}) = 0.025$, or 2.5% in percentage terms. The weighted prevalence average in 2000 over the twelve race-age groups (WNH, ages 18-49; WNH, ages 50-64; through to Other races, ages 65+) is 6.8%, with weights based on the US national population.

Table 1 shows wide regional differences: high rates in east south central and south Atlantic divisions contrast with low prevalence in the Mountain states. County poverty level also has a significant impact on prevalence after controlling for individual risk factors such as educational level and race. However, the latter are the predominant influences compared to contextual effects, with black ethnicity and elementary schooling being the most significant individual risk factors.

Table 2 shows estimates for the growth δ-parameters. The exponents of these parameters (plus the growth intercept) are growth rates over a decade for the category concerned, after partialling out the influence of other factors. Notable points are that the male excess apparent in Table 1 carries over into the growth component, so that the male-female prevalence gap is continuing to widen. Division effects on changing prevalence seem also mainly to preserve the "diabetes belt", though possibly extending that belt westward into states such as Texas. The highest regional growth rates are in the East South Central division (Kentucky, Tennessee, Alabama, Mississippi) and West South Central division (Texas,

Table 1: Influences on cross-sectional prevalence

		Mean	2.5%	97.5%	Relative risk compared to overall prevalence
	Intercept	-2.384	-2.396	-2.376	
Gender	Male	0.076	0.072	0.081	1.08
	Female	-0.076	-0.081	-0.072	0.93
Division	New England	-0.038	-0.056	-0.019	0.96
	Mid-Atlantic	0.062	0.048	0.075	1.06
	East North Central	0.092	0.081	0.101	1.10
	West North Central	-0.049	-0.061	-0.037	0.95
	South Atlantic	0.069	0.060	0.078	1.07
	East South Central	0.103	0.084	0.122	1.11
	West South Central	-0.020	-0.039	-0.001	0.98
	Mountain	-0.152	-0.167	-0.136	0.86
	Pacific	-0.065	-0.078	-0.052	0.94
County	Quartile 1 (Low poverty)	-0.104	-0.115	-0.094	0.90
poverty	Quartile 2	0.007	-0.003	0.014	1.01
	Quartile 3	0.050	0.039	0.060	1.05
	Quartile 4 (High poverty)	0.048	0.035	0.061	1.05
Age	18-49	-1.025	-1.032	-1.019	0.36
	50-64	0.347	0.341	0.353	1.41
	65+	0.678	0.673	0.685	1.97
Race	White N-H	-0.286	-0.296	-0.273	0.75
	Black N-H	0.266	0.245	0.285	1.30
	Hispanic	-0.020	-0.038	0.000	0.98
	Other	0.040	0.021	0.051	1.04
Education	Elementary Only	0.301	0.289	0.315	1.35
	High School Graduate	0.058	0.051	0.065	1.06
	Some College	0.009	0.001	0.018	1.01
	College Graduate	-0.368	-0.377	-0.361	0.69

Oklahoma, Arizona and Oklahoma), while growth is lowest in the Mid-Atlantic division.

By contrast, there is a realignment of race differentials through time, with black non-Hispanic prevalence rising relatively slowly (after controlling for other factors), and prevalence among other ethnic groups (mainly Asian Americans, and Native Americans) rising faster than in other race categories. Similarly, prevalence differentials linked to county poverty are not widening, and there seems to be a convergence, with higher than average growth in prevalence in low

Table 2: Influences on prevalence growth

		Mean	2.5%	97.5%
	Intercept	0.432	0.422	0.451
Gender	Male	0.0614	0.0548	0.0674
	Female	-0.0614	-0.0674	-0.0548
Division	New England	0.0033	-0.0263	0.0283
	Mid-Atlantic	-0.0608	-0.0805	-0.0409
	East North Central	-0.0402	-0.0556	-0.0249
	West North Central	-0.0144	-0.0329	0.0027
	South Atlantic	-0.0212	-0.0336	-0.0073
	East South Central	0.0596	0.0315	0.0853
	West South Central	0.1104	0.0823	0.1376
	Mountain	-0.0116	-0.0335	0.0084
	Pacific	-0.0251	-0.0453	-0.0085
County	Quartile 1 (Low poverty)	0.0219	0.0081	0.0376
poverty	Quartile 2	-0.0190	-0.0293	-0.0073
	Quartile 3	-0.0281	-0.0423	-0.0122
	Quartile 4 (High poverty)	0.0251	0.0076	0.0428
Age	18-49	0.1407	0.1314	0.1498
	50-64	-0.0741	-0.0835	-0.0642
	65+	-0.0667	-0.0755	-0.0582
Race	White N-H	-0.0179	-0.0378	-0.0024
	Black N-H	-0.0466	-0.0756	-0.0154
	Hispanic	0.0103	-0.0142	0.0373
	Other	0.0542	0.0384	0.0770
Education	Elementary Only	-0.0218	-0.0421	-0.0042
	High School Graduate	0.0306	0.0184	0.0400
	Some College	0.0233	0.0113	0.0345
	College Graduate	-0.0321	-0.0439	-0.0193

poverty as well as in high poverty areas. A convergence is also apparent in terms of education effects: the highest growth is among high school graduates and those with some college education as opposed to those with elementary education only – though growth is relatively low among full college graduates. This implies a widening gap between those with intermediate education qualifications as compared to college graduates. Finally apparent is a faster growth in prevalence among younger than older adults.

6. Prevalence Forecasts

Illustrative forecasts of county level prevalence - on the basis of these differential trends - are here made for 2011 (one year following the most recent survey) and for 2015. Figures 2 and 3 show male and female prevalence in 2011 (percent diagnosed prevalence with quintile break points), while Figures 4 and 5 show gender-specific prevalence in 2015. These maps confirm the continuing existence of a diabetes belt.

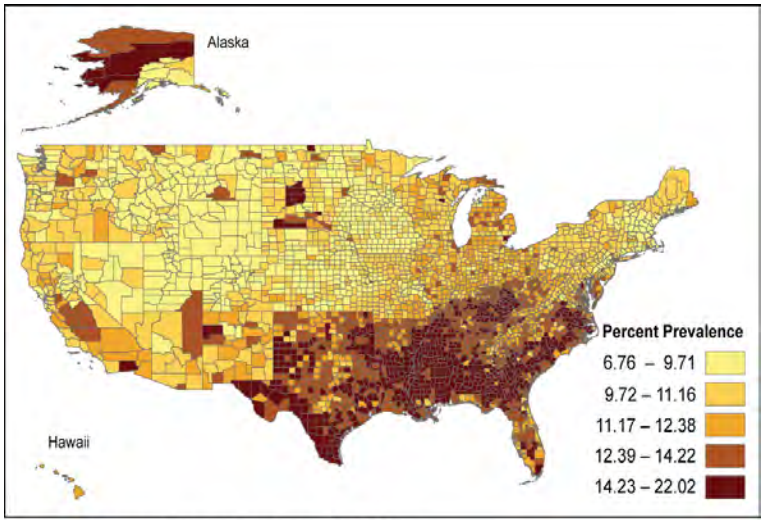

Figure 2: Diagnosed diabetes prevalence 2011 (Males)

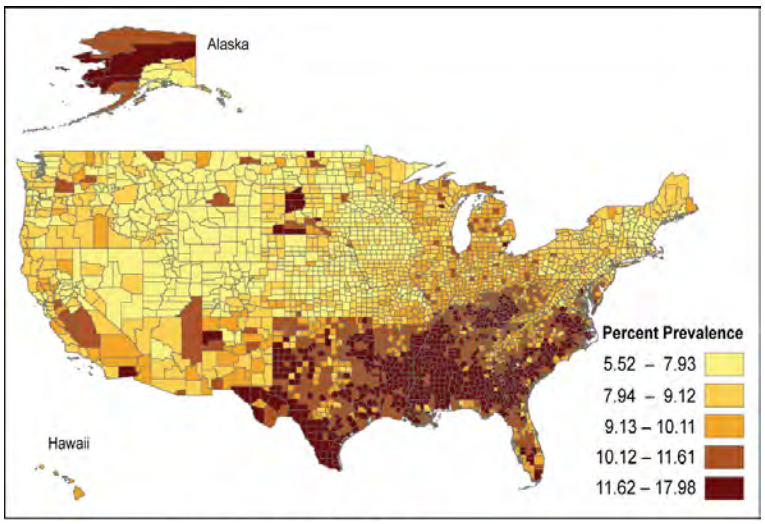

Figure 3: Diagnosed diabetes prevalence 2011 (Females)

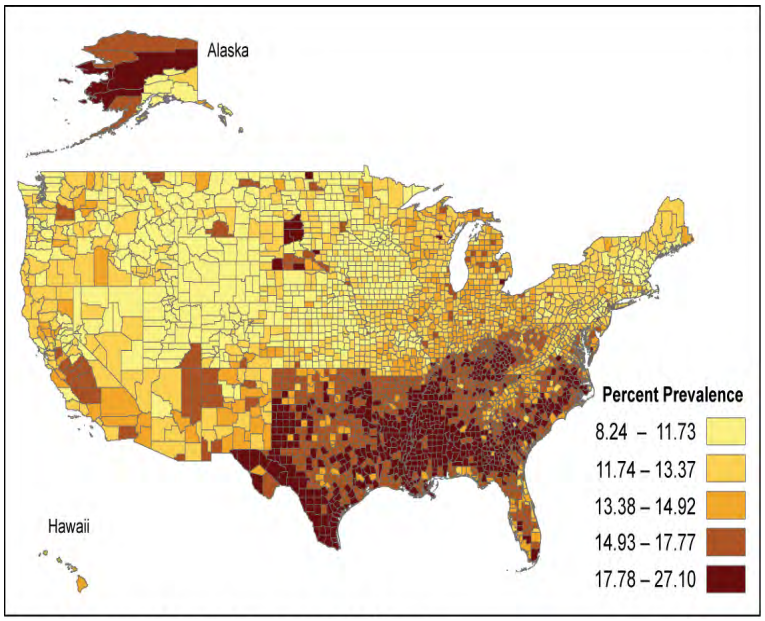

Figure 4: Diagnosed diabetes prevalence 2015 (Males)

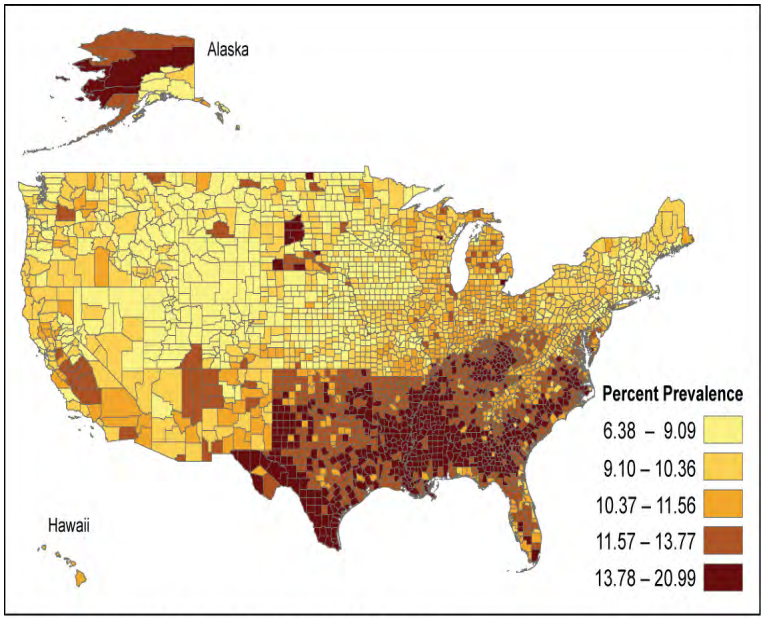

Figure 5: Diagnosed diabetes prevalence 2015 (Females)

Forecasts of diagnosed diabetes are likely to understate the disease burden, as the BRFSS survey relies on self-reported diagnosed diabetes and cannot mea-

sure prevalence of undiagnosed diabetes. To provide an estimate of total diabetes prevalence, results from the NHANES surveys for 1999-2008 were used. As recommended by the American Diabetes Association, a fasting plasma glucose level exceeding 126 mg/dL was used to ascertain undiagnosed diabetes among persons without previous diagnosed diabetes. Ratios of total to diagnosed prevalence were obtained for the 24 demographic groups used in the regression modelling: gender, age (18-49, 50-64 and 65+), and race (white non-Hispanic, black non-Hispanic, Hispanic, or other). These are used to scale prevalence rates defined by individual risk factors sex, age, and race (the US-wide probabilities π^{ind}_{gart}), and this scaling is maintained in county level estimates and forecasts. Under-diagnosis is higher among men than women, and among younger adults (ages 18-49).

Figures 6 and 7 accordingly show county level forecasts for each sex of total diabetes prevalence in 2011. These are summarised in Table 3 which contrasts 2011 forecast levels of total diabetes prevalence according to the five highest ranked states (for male prevalence) and the five lowest ranked states. There is a two-fold difference in total prevalence between the highest and lowest ranked states, widening to a three-fold difference (26.1% vs 8.9% for males, 20% vs 6.6% for females) between the five highest ranked and five lowest ranked counties. The three highest ranked counties are in Mississippi (Jefferson County, Humphreys County and Holmes County).

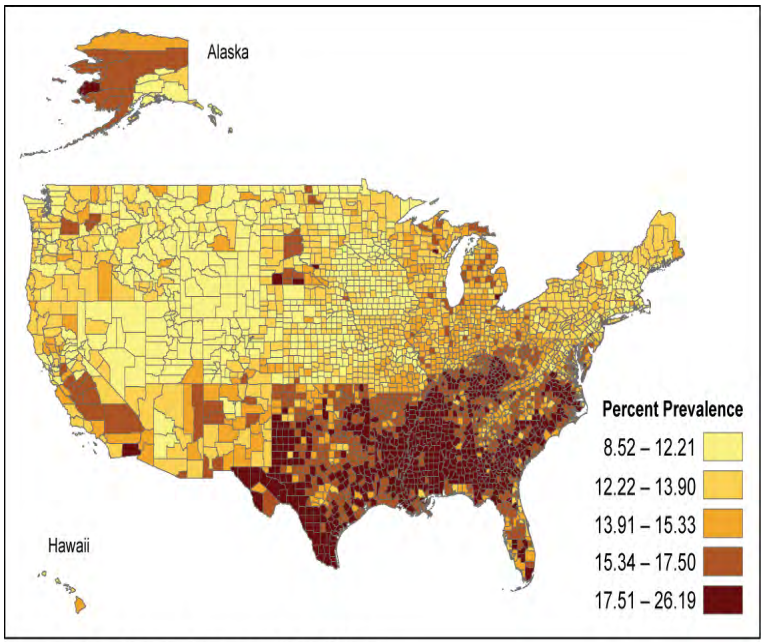

Figure 6: Total diabetes prevalence 2011 (Males)

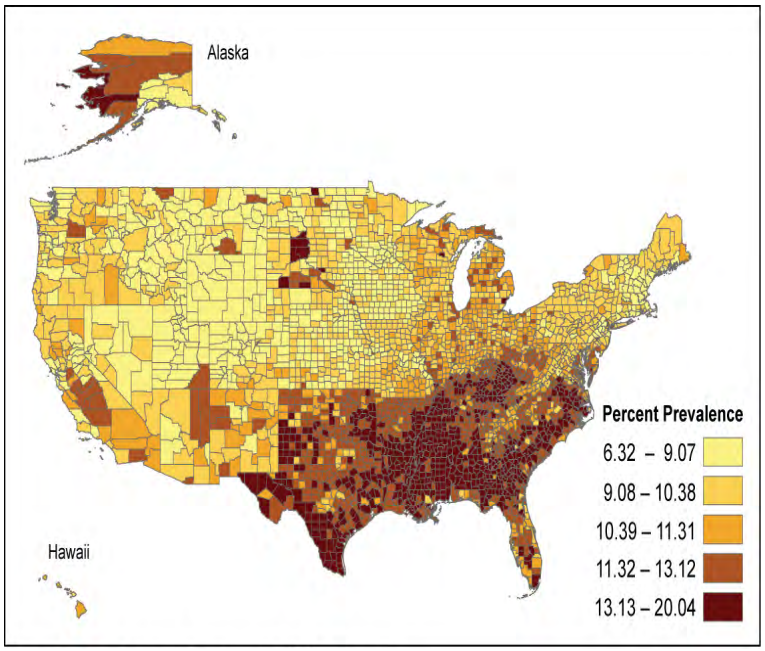

Figure 7: Total diabetes prevalence 2011 (Females)

Table 3: Total adult diabetes prevalence (%), including undiagnosed, highest and lowest age standardised rates by state 2011

Highest ranked	M	F
Mississippi	20.4	15.4
Alabama	18.7	14 1
Louisiana	18.2	13.7
Tennessee	17.8	13.3
Texas	17.5	12.9
Lowest ranked	M	F
New Hampshire	11.2	8.4
Colorado	11.1	8.2
Montana	11.0	8.2
Utah	10.8	8.0
Wyoming	10.6	7.8

7. Conclusions

While nationwide forecasts of diabetes prevalence in the US have been produced, so far disaggregated forecasts are not available. However, geographically disaggregated forecasts are important for planning public health interventions,

as there are wide variations in diabetes prevalence between small areas (counties and below) and regions of the US (e.g., Barker *et al.*, 2011; Ford *et al.*, 2005).

To enable geographically disaggregated forecasts, a weighted binomial regression is applied to BRFSS totals of diagnosed diabetes cases, and populations at risk. Observations are defined by survey year (2000 to 2010), individual risk/protective attributes, and residence category. The model includes a cross sectional component together with a linear growth path in decadal time units, and forecasts are obtained by extending the growth path to future years.

Significant influences on cross-sectional prevalence variations during 2000-2010 include black ethnicity, limited education, and living in a high poverty county. Prevalence growth is highest among younger adults, among males, and among those with high school education. Allowing for undiagnosed prevalence (e.g., Figures 6 and 7) emphasizes the gender gap in diabetes prevalence trends, and also regional shifts with a widening of the diabetes belt.

Trends in diabetes are likely to be linked to changes in risk factors such as obesity, and the methods outlined here for geographically disaggregated forecasts can be applied also to such co-morbid conditions for diabetes.

Acknowledgements

The author acknowledges the support of the National Minority Quality Forum.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition. Wiley, New York.

Barker, L. E., Kirtland, K. A., Gregg, E. W., Geiss, L. S. and Thompson, T. J. (2011). Geographic distribution of diagnosed diabetes in the U.S.: a diabetes belt. *American Journal of Preventive Medicine* **40**, 434-439.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1-20.

Blizzard, C. L. and Hosmer, D. W. (2006). Parameter estimation and goodness-of-fit in log binomial regression. *Biometrical Journal* **48**, 5-22.

Boyle, J. P., Thompson, T. J., Gregg, E. W., Barker, L. E. and Williamson, D. F. (2010). Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence. *Population Health Metrics* **8**, 29.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434-455.

Cowie, C. C., Rust, K. F., Byrd-Holt, D. D., Eberhardt, M. S., Flegal, K. M., Engelgau, M. M., Saydah, S. H., Williams, D. E., Geiss, L. S. and Gregg, E. W. (2006). Prevalence of diabetes and impaired fasting glucose in adults in the U.S. population: National Health and Nutrition Examination Survey 1999-2002. *Diabetes Care* **29**, 1263-1268.

Ford, E. S., Mokdad, A. H., Giles, W. H., Galuska, D. A. and Serdula, M. K. (2005). Geographic variation in the prevalence of obesity, diabetes, and obesity-related behaviors. *Obesity Research* **13**, 118-122.

Gelfand, A. E. (1996). Model determination using sampling based methods. In *Markov Chain Monte Carlo in Practice* (Edited by W. Gilks, S. Richardson and D. Spiegelhalter, Markov Chain Monte Carlo in Practice), 145-161. Chapman & Hall, Boca Raton, Florida.

Huang, E. S., Basu, A., O'Grady, M. and Capretta, J. C. (2009). Projecting the future diabetes population size and related costs for the U.S. *Diabetes Care* **32**, 2225-2229.

Marshall, E. C. and Spiegelhalter, D. J. (2007). Simulation-based tests for divergent behaviour in hierarchical models. *Bayesian Analysis* **2**, 409-444.

Maty, S. C., Everson-Rose, S. A., Haan, M. N., Raghunathan, T. E. and Kaplan, G. A. (2005). Education, income, occupation, and the 34-year incidence (1965-99) of Type 2 diabetes in the Alameda County Study. *International Journal of Epidemiology* **34**, 1274-1281.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Nelson, D. E., Engelgau, M. M., Vinicor, F. and Marks, J. S. (2000). Diabetes trends in the U.S.: 1990-1998. *Diabetes Care* **23**, 1278-1283.

Monnat, S. M. and Pickett, B. C. (2011). Rural/urban differences in self-rated health: examining the roles of county size and metropolitan adjacency. *Health and Place* **17**, 311-319.

National Institute of Diabetes and Digestive and Kidney Diseases. (2011). http://diabetes.niddk.nih.gov/dm/pubs/statistics/#Racial. Accessed July 2.

Smith, J. (2007). Diabetes and the rise of the SES health gradient. National Bureau of Economic Research Working Paper 12905. Cambridge, Massachusetts.

Received December 28, 2011; accepted April 30, 2012.

Peter Congdon
School of Geography
Queen Mary University of London
Mile End Road, London, E1 4NS, UK
p.congdon@qmul.ac.uk