Defining simulated populations

This is an attempt at a fairly precise definition of what we want to do (at least as a first pass) in simulating populations to be sampled. I will try to indicate directions for future expansion ...

Populations

It may eventually make sense to be able to combine multiple separate populations (communities? municipalities?) of different sizes and characteristics into a single sample, but we're not doing that in the current iteration.

Tile characteristics

We had decided it would probably make sense for purposes of efficiency/scalability to break a given sample up into "tiles", within which there was some egree of homogeneity. For now the tiles are going to be considered rectangular, but this could easily be generalized – the only real constraint is that they be non-overlapping, and probably contiguous.In principle there could be separate coordinate systems within each tile, although we would need some functionality for converting to absolute coordinates.

- We will consider an `n_tile_x` × `n_tile_y` rectangular grid, with size `L_tile`. Thus the centroids of the tiles are at `((i+1/2)*L_tile,(j+1/2)*L_tile)` for `i,j` equal to `0..(n_tile-1)`. (It will probably be best to use realistic distance units.) For our example, let's say `n_tile_x=n_tile_y`=10, `L_tile`=3.5 km (this gives us 1225 km^2, a little bigger than Hamilton)
- We can vary characteristics spatially across tiles; if the variation is spatially structured, we will associate it with the centroids (*or other characteristic location*) of the tiles. Several reasonable ways to generate characteristics would be:
- Simple random sampling (i.e. spatially uncorrelated; each tile gets an independent sample, e.g. Normal with specified mean and standard deviation)
- Linear or quadratic trend surface (i.e. the value of centroid $\{x, y\}$ is $b_{00}+b_{01}x+b_{10}y+b_{20}x^2+b_{02}y^2+b_{11}xy$)
- A specified Gaussian random field (perhaps overkill)
- for now, let's suppose that income has a linear trend and average population density is randomly distributed among tiles. The units of income are more or less arbitrary: let's say we use `inc=0.75+x*1/(4*35)+y/(4*35)` so the income ranges from 0.75 at `{0,0}` to 1.25 at `{1,1}`. We want an average population density of 400/km^2 (again, using Hamilton as an example), so let's say the population density of a tile is chosen from `Normal(mean=400,sd=50)`.

Within-tile spatial characteristics

We may want to generate spatial variation within tiles, especially of disease risk. As discussed previously there are a lot of ways to do this ("pocketing", defined by picking point locations and computing a risk based on a *kernel* (a declining function of distance from the disease foci), or some sort of Gaussian random field). These could be associated with the tiles, or with the households within the tiles.

Buildings

We haven't decided on the distribution of buildings within tiles. They could be regular, random, or chosen from some point process (e.g. Strauss process). We need to pick the number of buildings per tile to match the desired population density for the tile: number of buildings = `popdens/(avg_households_per_building*avg_indiv_per_household)`. For now we will assume 1

household per building (later we may want to relax this, and in particular say that the numbers of households per building are correlated with overall population density ...) and 5 individuals per household.

For now let's make the building distribution random within tiles (unless we want to try the exercise of patching in a Strauss-process algorithm from R or somewhere else ...)

Households

Households inherit spatial locations from their buildings.

Now we need to generate individuals within households.

- **Household characteristics**: we might assign some characteristics (income, ethnicity, etc.) at the household level; these would be determined by some random variation around the tile-level averages.

- Let's say that household income is a Normal deviate from the tile average with a standard deviation of 0.25 (remember that the overall income range is (0.75-1.25), so there's a fair amount of within-tile variation).
- Let's also assign a *latent variable* for disease risk within the household – this accounts for either contagion within households, or the effects of all the other among-household differences that aren't captured by the particular covariates we decided to simulate. Let's make this a standard normal `N(0,1)` variable (we can decide later how to scale its impact)
- **Household size**: We want a positive distribution (no empty households!). We could use a truncated Poisson distribution, but it's a little simpler (if we are aiming for a mean of λ individuals per household) to use $1 + \text{Poisson}(\lambda - 1)$.
- **Sex/age distribution**: HH of size 1 are randomly male vs female adults. HH of size 2 are assumed to be heterosexual couples (1 M adult, 1 F adult). HH of size >2 are adults plus children (randomly assigned sexes). For now we won't try to simulate age distributions, just categorize individuals as adults or children.
- **Other individual-level characteristics**: we could also assign individual-level latent disease risk scores, to increase the within-household variation, but let's not bother for now.

Assigning disease status

Now we assign disease status to individuals. We are going to do this in a standard generalized-linear-model way: that is, we are going to construct a linear function of the various contributing factors (income, spatial disease effect, household disease effect, individual disease effect, age/adult vs. child). In order to make this into a probability we then logistic-transform the linear disease score (called a *linear predictor* in statistical contexts), and choose a Bernoulli random variable based on this score.

The most comprehensible way to do this will probably be to * standardize all of the predictors (income, household latent variable, etc.) to have mean zero and standard deviation 1 (the latent variables are already scaled this way); * scale each variable by an "importance parameter" β and add them together; * add an intercept to set the desired median prevalence.

For example, we might decide that income and household latent variable have equal weights ($\beta_1 = \beta_2 = 1$), and that we want a median prevalence of 10%, so the intercept $\beta_0 = \text{logit}(0.1) = -2.2$.

Thus the general form of the disease incidence model is

$$\eta = \beta_0 + \sum_i (x_i/\bar{x}_i)/\sigma$$

$$y_i \sim \text{Bernoulli}(\text{logistic}(\eta))$$