USDOS 2.1 Shipment Generation Process

Stefan Sellman

December 18, 2015

1 Number of shipments

The number of shipments that originate from a premises is drawn from a Poisson distribution (eq. 1). The Poisson distribution gives the number of events that occur within a given period of time if the events occur at a known average rate. The rate (λ), in this case, is the average daily number of shipments originating from the state that the premises is in, divided by the number of premises in that state (eq. 2). Premises are not weighted in any way within the state.

$$n_i \sim Pois(\lambda_s) \tag{1}$$

where n_i is the number of shipments for premises i, s is the state of farm i and λ_S (mean of Poisson distribution) is:

$$\lambda_s = \frac{1}{365}\frac{n_s}{N_s} \tag{2}$$

where n_s is the total number of shipments originating from state s in a year and N_S is the number of premises in s. n_s is currently input to the program through the file temp_n_<animal type>.csv[1].

2 Destination county probabilities

The first step in deciding the destination of the shipment is to pick which county it goes to. The probabilities for sending to each other county or the county itself is a function of the distance between them (the shipment distance kernel) and an animal type specific shipment weight (eq 3).

$$P_s(i \rightarrow j) = \frac{w_{j,t} e^{-(\frac{d_{ij}}{a})^b}}{\sum_{k=1}^{C} w_{k,t} e^{-(\frac{d_{ik}}{a})^b}} \tag{3}$$

[1] A file with one row of floating point values, separated by commas, describing the a, b, or n for each county or state respectively.

i is the origin county, j is the destination county, w is the county specific weight and C is the total number of counties. a and b are specific kernel parameters, estimated by USAMM for each county and describe the shape of the distance kernel. $w_{j,t}$ and $w_{k,t}$ are the county and animal type specific shipment weights, based on historical flow data. Values for w are input to the program through the file fips_weights.txt. a and b are input to the program through the files temp_a_<animal type>.csv[1] and temp_b_<animal type>.csv[1]. The sum of all probabilities for county i are all normalized to one.

County-to-county probabilities are calculated as needed the first time a county has a shipment that originates from a premises within it. They are then reused until the county goes out of scope (is deleted; this currently causes a sneaky bug if different a's & b's are used each replicate since the counties are just reused every replicate and doesn't get deleted and created anew). The probabilities are stored in an advanced data type called alias table [1]. An alias table can be used to represent a discrete probability distribution and is both memory efficient ($\Theta(n)$), quick to initiate ($\Theta(n)$) and efficient at generating random draws from the distribution ($\Theta(1)$).

The current way of assigning shipments by animal type once a premises has been determined to send something is to first generate the destination county and then, from the subset of premises in the destination county that have the same animal type as the sending premises, pick one destination at random. No weighting whatsoever. If there are no premises with matching animal type, then a new destination county is generated, until a matching destination premises is found. This last part should really never happen since if the county doesn't have any premises of a certain type, then it's inflow parameter (w) should equal 0 and therefore the probability of sending there will also equal 0. Depending on how the landscape is generated and how premises types are assigned to the generated population, I guess this issue could arise by random chance (no premises of a certain type generated within a county that we know have an inflow for that type).

3 Distance binning

Because the shipment parameters currently used were created using binned distances. This was originally done to improve the speed of calculating the parameters since the distance kernel would only have to be evaluated for a finite set of discrete distances. For compatibility with the way that the parameters were created there exists a way to force the distances in USDOS shipment generation into bins as well. The distance is just converted to the correct distance bin and then the county-to-county probability is calculated for that distance bin. The binning only leads to an improvement in processing speed if the kernel value for that specific bin is saved for later use *and* that kernel value gets reused a reasonable number of times. The binned kernel value in USDOS is currently not saved and the rationale behind that is that in relation to USAMM (where binning is more or less essential) the number of times the kernel will need to be evaluated are few owing to the small number of shipments that are generated. If we would

find ourselves in a situation where we have significantly more shipments generated we should consider using "true" binning. Until then I propose we stick to this "false" binning since the extra overhead would probably be more than the speed gained. The option to use "false" binning or not is currently hard-coded. It can easily be changed with a flag, but requires a recompilation. Should it be an option in the config file instead? That would remove the need for recompilation and it could be turned off and on by the user on a simulation-to-simulation basis.

4 County probabilities with squared distance

There are currently two ways to calculate the same shipment probability coded into the program. One is 3, the other is a rearranged form of the kernel part of 3 that avoids using the square root operation to calculate the euclidean distance. Instead of the square root it uses two extra multiplications which is marginally faster.

$$e^{-\left(\frac{d_{ij}}{a}\right)^b} = e^{-\left(\frac{\sqrt{\Delta x^2+\Delta y^2}}{a}\right)^b} = e^{-\left(\sqrt{\frac{\Delta x^2+\Delta y^2}{a^2}}\right)^b} = e^{-\left(\frac{\Delta x^2+\Delta y^2}{a^2}\right)^{0.5b}} \quad (4)$$

The variables are the same as eq, 3 but Δx and Δy are the differences in x and y between the counties centroids (euclidean distance without square root).

References

[1] Michael D. Vose A Linear Algorithm For Generating Random Numbers With a Given Distribution IEEE Transactions of software engineering Vol. 17, No. 9. 1991