Designing Sustainable Landscapes: Probability of Development

A project of the University of Massachusetts Landscape Ecology Lab

Principals:

- Kevin McGarigal, Professor
- Ethan Plunkett, Research Associate
- Brad Compton, Research Associate
- Bill DeLuca, Research Associate
- Joanna Grand, Research Associate

With support from:

- North Atlantic Landscape Conservation Cooperative (US Fish and Wildlife Service, Northeast Region)
- Northeast Climate Science Center (USGS)
- University of Massachusetts, Amherst

Report date: 20 April 2018

Reference:

McGarigal K, Plunkett EB, Compton BW, DeLuca WV, and Grand J. 2017. Designing sustainable landscapes: probability of development. Report to the North Atlantic Conservation Cooperative, US Fish and Wildlife Service, Northeast Region.

General description

The integrated probability of development (probDevelop) is derived from an extraordinarily complex urban growth model described in detail in the technical document on urban growth (McGarigal et al 2017). The urban growth model is one of the major landscape change drivers in our Landscape Change, Assessment and Design (LCAD) model, in which it functions to simulate the stochastic growth of low-, moderate- and high-intensity development during each 10-year timestep of a 70-year simulation between 2010-2080. Because the urban growth model simulates the spatial footprint of development as a stochastic process, the development that occurs in any one simulation is merely a single stochastic realization of what could happen. Thus, any single urban growth realization is not particularly useful in landscape design. Instead, we developed this product to represent the integrated probability of development occurring between 2010-2080, which accounts for the type (low-, medium-, and high-intensity), amount and spatial pattern of development. This index represents the probability of development integrated across all of the possible development transitions occurring sometime between 2010 and 2080 at the 30 m cell level (**Fig. 1**).

Briefly, in the urban growth model the projected amount of future development in an area is downscaled from county level forecasts based on a U.S. Forest Service 2010 Resources Planning Act (RPA) assessment (Wear 2011) to individual application "panes" ~5 km on a side. Within an application pane the type (i.e., new low, new medium, new high, low to medium, low to high, and medium to high) and spatial pattern of development at the 30-m cell level is based on statistical models of historical development and is influenced by factors such as geophysical conditions (e.g., slope, intensity of open water) and proximity and intensity of roads and urban development. Ultimately, each 30-m cell ends up with a probability of each type of development that reflects the total projected demand for development in the application pane and the

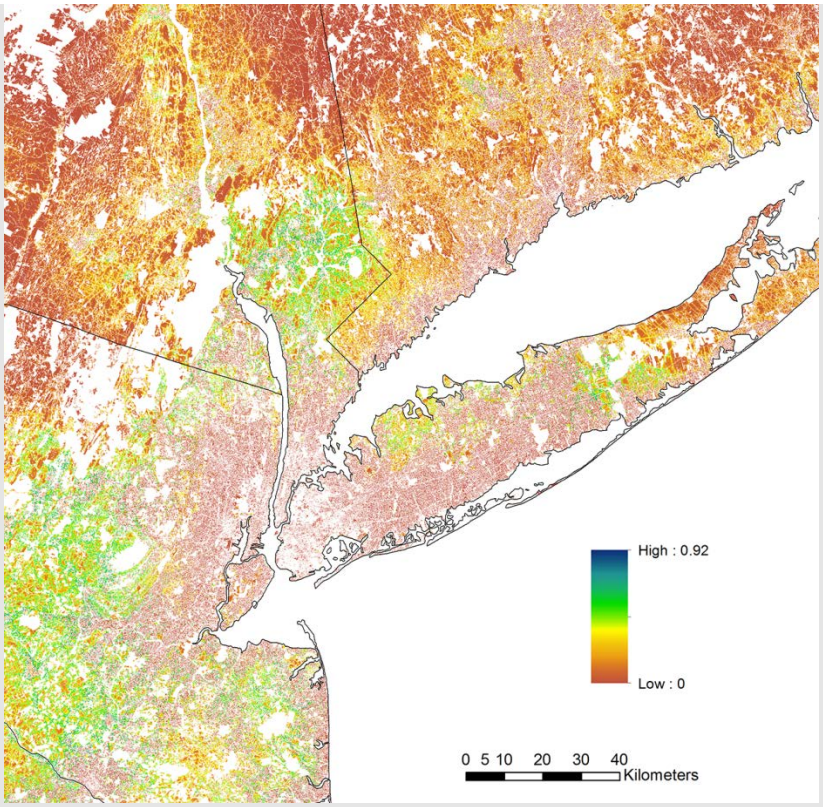

Figure 1. Integrated probability of development occurring between 2010-2080 for the area in the vicinity of New York. Areas shown in white are unbuildable (e.g., water, barren, secured).

relative likelihood of that type of development occurring on that cell given its spatial context. For this product we simply compute the cumulative probability of a cell being developed, weighted by the type of development, over the 70-year period. The end result is a seamless and continuous representation of the integrated probability of development occurring between 2010-2080 at the 30 m cell level.

Use and interpretation of this layer

This product can be used in combination with any of the other landscape conservation design (LCD) products that reveal places of high ecological value to indicate places of ecological value that are at risk of development and thus may warrant land protection. This product also can be used to identify places at risk of future development independent of designated core areas and any formal LCD. Note, although this index is a true probability, it is perhaps best used in a relative manner to compare values from one location to another. Its use should be guided by the following considerations:

- It is important to acknowledge that the integrated probability of development surface was derived from a model, and thus subject to the limitations of any model due to incomplete and imperfect data, and a limited understanding of the phenomenon being represented. In particular, the GIS data upon which this product was built are imperfect; they contain errors of both omission and commission. This is especially true for the National Land Cover Dataset (NLCD) from which development is mapped and the probability of development is modeled. Consequently, there will be many places where the model gets it wrong, not necessarily because the model itself is wrong, but rather because the input data are wrong. Thus, the probability of development surface should be used and interpreted with caution and an appreciation for the limits of the available data and models. In particular, this surface is probably best used as a general indication of where development is likely to occur, but at the cell level it is not expected to be highly reliable. However, getting it wrong in some places should not undermine the utility of the product as a whole. As long as the model gets it right most of the time, it still should have great utility.
- It is important to recognize that the integrated probability of development is highest near existing roads, largely because our urban growth model does not attempt to predict the building of new roads and the development associated with them. Because proximity to roads is an important and dominant predictor of development at the 30-m cell level in our model, our integrated probability of development surface is going to be heavily biased towards existing roads. This means that we don't do a very good job of predicting where a subdivision might get developed in the future.
- This product is combined with the HUC6 local conductance index to create the HUC6 local vulnerability index (see vulnerability document, McGarigal et al 2017), a core-independent measure that reflects the likelihood of development occurring in places with high local conductance (**Fig. 2**). Cells that confer high local conductivity at the scale of one to a few kilometers that also have high probability of development are most vulnerable and thus could represent priorities for land protection.
- This product is combined with the HUC6 regional conductance index (see conductance document, McGarigal et al 2017) to create the HUC6 regional

vulnerability index (see vulnerability document, McGarigal et al 2017), a core-dependent measure that reflects the likelihood of development occurring in places that confer connectivity between the designated terrestrial cores (see terrestrial core area network document, McGarigal et al 2017) (**Fig. 2**). Cells with relatively high regional conductance between terrestrial cores where the flow is concentrated in narrow "corridors" (and thus irreplaceable) and where the probability of development is relatively high are most vulnerable and thus could represent priorities for land protection.

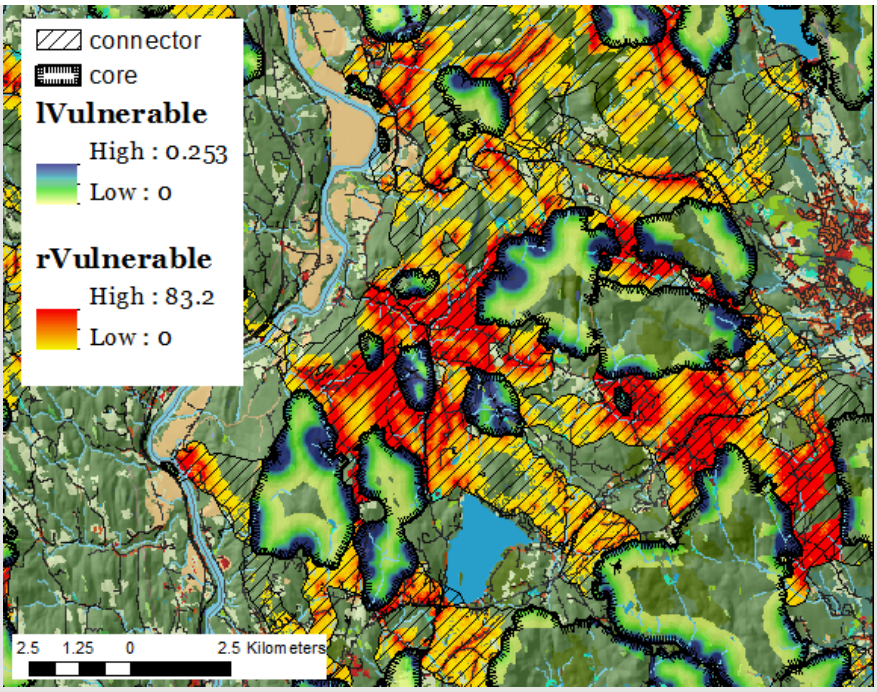

Figure 2. Vulnerability of conductance to future development depicted by a combination of the local vulnerability index (lVulnerability) within terrestrial core areas and the regional vulnerability index (rVulnerability) within connectors. Areas in dark blue within cores and dark red within connectors have a high risk of future development.

Derivation of this layer

The derivation of the integrated probability of development layer is extraordinarily complex, as described in detail in the technical document on urban growth (McGarigal et al 2017). To fully understand the derivation of this layer, it is necessary to understand the urban growth model from which it is derived. Here, we describe a highly abbreviated version of the process.

1. Training data

The urban growth model utilizes historical training data to characterize urban growth patterns for six different transition types:

transition 1 = undeveloped to low-intensity development;
transition 2 = undeveloped to medium-intensity development;
transition 3 = undeveloped to high-intensity development;
transition 4 = low-intensity development to medium-intensity development;
transition 5 = low-intensity development to high-intensity development; and

transition 6 = medium-intensity development to high-intensity development.

The training data were taken from a subset of the Northeast, specifically Maine and Massachusetts, and the Chesapeake Bay. The primary training data source for Maine and Massachusetts was the National Oceanic and Atmospheric Administration (NOAA) Coastal Services Center Coastal Change Analysis Program (C-CAP) data. These data were available for different time periods for many coastal areas within the U.S. However, not all inland areas were available. Much of the Northeast was available in grid format for the 1996 and 2006 timesteps: http://www.csc.noaa.gov/crs/lca/northeast.html. We also utilized Chesapeake Bay Watershed Landcover Data Series (CBLCD) available from 1984 and 2006: ftp://ftp.chesapeakebay.net/Gis/CBLCD_Series/, which was based partly on the C-CAP data as well.

We divided each of the three training regions into non-overlapping training "windows" ~15 km on a side. For each training window, we computed the Gaussian kernel (12.8 km bandwidth) intensity of development and the intensity of open water, and converted these values to z-scores (i.e., mean=0 and standard deviation=1). Thus, each training window occupied a position in a standardized two-dimensional state space defined by the intensity of development and open water. We located 16 uniformly distributed "model points" or locations in this state space. For each of these model points, we fit separate binary logistic regression models for each of the six transition types to a set of training points (i.e., cells of the corresponding transition type a minimum distance apart matched with an equal number of randomly selected "available" cells) from the training windows located within 1, 1.5 or 2 standard deviations from the model point in the model state space, ensuring a minimum of 200 training points (half of which experienced that transition and half of which were available points) were included in the model. If we could not meet the minimum sample size of training points, we dropped the model point from consideration. The predictor variables in the logistic regression models included:

- Gaussian kernel (bandwidths=100, 800 and 3,200 m) intensity of open water;
- Gaussian kernel (bandwidth=800 m) intensity of primary and secondary roads;
- Gaussian kernel (bandwidth=800 m) intensity of all roads except motorways;
- Gaussian kernel (bandwidth=3,200 m) intensity of all roads except motorways;
- Gaussian kernel (bandwidths=400 and 3,200 m) intensity of weighted development (weights for low-intensity development=1, medium-intensity development=2, high-intensity development=3; NA on all cells not eligible for development);
- Transformed slope based on a univariate logistic regression model; and
- Transformed distance to the nearest road (excluding motorways) based on a univariate logistic regression model.

Ultimately, we ended up with a logistic regression model to predict each of the six transition types for each of 12-16 model points, depending on transition type, uniformly distributed throughout the state space defined by the intensity of development and open water. After preliminary evaluation of the results we noted that our logistic regression models were not forcing enough of the new development (transitions 1-3) to be close to roads. This was due to bias in the training data resulting from geo-processing to eliminate

the confounding of roads and development in the C-CAP and CBLCD data. To adjust for this bias, we multiplied the logistic functions for transitions 1-3 by a negative exponential function of distance to nearest road. The latter was fit to the distribution of new developments (transitions 1-3 pooled) observed in a hindcast dataset we developed to validate the urban growth model (described in the technical documentation).

In addition, for each training window we also computed the historical distribution of transition types (i.e., the proportion of total transitions comprised of each of the six transition types), the distribution of observed sizes of disjunct development patches, and the total amount (in cells) of historical development - the "match amount", for use in the simulation (see below).

2. Urban growth model

The urban growth model consists of several interacting components. The basis for the "current", or initial, land-use condition in the LCAD simulation (set to be roughly the year 2010) is the set of developed landcover classes in DSLland (see DSLland document, McGarigal et al 2017), including low-, medium- and high-intensity development derived from the 2011 National Landcover Dataset (NLCD).

For purposes of the urban growth model, we subdivide the entire Northeast region into non-overlapping square application "panes" ~5 km on a side, each of which is embedded as the central pane within a square application "window" consisting of 3x3 panes (~15 km on a side). Given this spatial template, the urban growth model is implemented as follows:

1) *Demand* — To begin, we establish the demand for additional cells of urban land-uses (including low-, medium-, or high-intensity development) at each 10-year timestep from 2010 to 2080. The demand dictates the overall *amount* (in cells) of urban land-uses to allocate throughout the area of interest in each timestep. The demand for urban growth at each future timestep is based on county-level forecasts derived for a U.S. Forest Service 2010 Resources Planning Act (RPA) assessment (Wear 2011). We aggregate these county-level forecasts into census Core Base Statistical Areas (CBSAs) where they exist, and otherwise retain the forecasts at the county level for those counties not in CBSAs. We convert the RPA forecasts given in absolute area to development rates, computed as the proportion of land area developed, and then convert this to an absolute demand (in cells) for each CBSA or county by multiplying the forecasted rate of development by the count of land cells within the CBSA or county.
2) *Matching* — Next, to allocate the total demand (in cells) within each CBSA or county to each application pane at each timestep, each application window is matched to the three most similar training windows based on geographic proximity and four landscape metrics: Gaussian kernel (12.8 km bandwidth) intensity of 1) development, 2) roads, and 3) open water, and the density of roads within the window. Note, these metrics were selected using a matching algorithm developed a priori during the training phase using the training data from all three training regions.
3) *Allocation* — Once each application window is matched to three training windows, we allocate the total demand (in cells) within the corresponding CBSA or county for the current timestep to each application pane. To do this, we calculate the total amount of

historical development (the "match amount") in the three training windows. This match amount is subsequently adjusted to reflect the proportion of the application window that is buildable and the proportion of the buildable in the window that is in the central pane. We also adjust this match amount as necessary to ensure that no more than 14% of the buildable cells for transitions 1-3 (i.e., available for new development) in the pane are built in any one decade. This development threshold was based on the 99th percentile of the corresponding distribution observed in a hindcast dataset we developed to validate the urban growth model (described in the technical documentation). The result is an interim measure of the amount to allocate to each application pane that reflects the historical distribution among panes having a similar landscape context. Lastly, the absolute demand (in cells) for each application pane in the current timestep is computed by dividing the pane's interim match amount by the total interim match amount across all panes in the corresponding CBSA or county. In this manner, the total demand (in cells) for each CBSA or county is allocated among application panes such that the more historical development that occurred in the matched training windows the higher proportion of the future demand is assigned to the application pane.

Next, the demand (in cells) in each application pane for the current timestep is allocated among the six transition types based on the historical distribution in the matched training windows, with the sum of the first three transitions (i.e., undeveloped to low-, medium- or high intensity developed) made to match the total allocation to the pane and the ratio among all six transitions made to match the historical ratios in the matched training windows.

4) *Suitability* — For each transition type we create an inverse distance-weighted average logistic regression model based on the distance between the application window and each model point in the two-dimensional model state-space described above. Next, we use these weighted-average models to compute the relative probability (i.e., suitability) of each transition type for each 30 m cell in the application pane. Note, because the Gaussian kernel (12.8 km bandwidth) intensity of development surface is changing over time due to urban growth, the position of each application window in the two-dimensional model state space is shifting over time as well. Consequently, the patterns of urban growth in an application window will shift over time and become more like the patterns characteristic of increasingly urbanized windows

5) *Disturbance patches* — Given the demand (in cells) for each transition type allocated to each application pane for the current timestep and the corresponding suitability surfaces, we simulate actual development for each transition type, as follows:

   a) Randomly select a cell to initiate the disturbance based on the relative probability (i.e., suitability) surface fit for that transition;

   b) randomly draw a patch size from the observed distribution of patch sizes in the three matching training windows for the corresponding transition type.

   c) spread outward from the initiation cell with a resistant kernel, where resistance is based on a multiple of the complement of the corresponding probability of transition for each neighboring cell, until the randomly selected patch size is met, allowing patches to extend across the boundaries of the focal application pane.

d) repeat the process above, building development patches sequentially until the total allocation of cells for the transition type is exhausted in the application pane.

Urban growth scenarios.—Urban growth scenarios can be implemented in two ways. First, the overall amount of land that is developed can be modified from the baseline demand computed above; e.g., to increase or decrease the amount of development relative to the RPA forecasted amount. For this product, we elected to utilize the unadjusted RPA projections. Second, the other factor that can be adjusted is the 'sprawl dial'. This dial operates across application windows, determining how contagious (compact vs. 'sprawlly') growth patterns will be at the broad scale, compared to historic trends. For this product, we elected to keep the sprawl dial at neutral, emulating the historical sprawl patterns.

In summary, the urban growth model acts as a disturbance process on the landscape, realizing development at the 30 m cell and patch level in each 5 km pane at each 10-year timestep until the allocated number of cells to be disturbed have been exhausted. The types of disturbance transitions (e.g., undeveloped to low intensity development, or medium- to high-intensity development) are allocated proportionately to that observed historically in the most similar training windows. The patterns of development for each transition type are modeled to reflect the historical patterns that occurred in landscapes having a similar landscape context. The sizes of patches developed are chosen to reflect the distribution observed historically in the most similar training windows. At the end of each 10-year timestep, once growth is realized, the resulting urban grid is fed back into the beginning of the process for the next timestep. Importantly, at each timestep each application window is matched to three new training windows and the weights assigned to each training model are recalculated . In this way, the model is non-stationary across space and time; as a window becomes more urbanized in the future, its growth patterns change to match the way more urbanized windows grew historically, but all subject to the projected demand for growth at the CBSA or county level.

3. Integrated probability of development

As described above, the urban growth model simulates urban development over time as a stochastic process. The output is a new human land-use layer (depicting low-, medium-, and high-intensity development) for each timestep, which represents a single stochastic realization of the urban growth process. While this is useful for landscape change simulation, it has limited utility as a single product to inform landscape conservation design. For the latter, we derived the integrated probability of development layer, which is a seamless and continuous representation of the cumulative probability of a cell being developed over the 70-year period (2010-2080) weighted by the type of development transition (**Fig. 1**).

One way to think about the integrated probability of development is as follows. It is roughly equivalent to running the urban growth under the baseline scenario thousands of times and computing the proportion of the reps in which each cell underwent each transition type sometime during the 70-year simulation. It is not exactly equivalent to this, however, because of the non-stationarity of the model. Specifically, in the urban growth model the stochastic development at each timestep changes the development footprint in each application pane, which influences subsequent development in that pane during the next timestep due to the unique matching algorithm. Thus, the exact trajectory of development

that occurs evolves during the simulation in response to the changing landscape. Averaging the results across thousands of replicate simulations would capture this non-stationarity and produce accurate results. However, it is impractical to run thousands of replicate simulations. Therefore, we are forced to treat the development patterns during the first timestep as effectively stationary and simply adjust the probabilities to account for the cumulative amount of development projected for the 70-year period. Thus, although the integrated probability of development is not exactly correct, we deemed it sufficiently correct to warrant its use in the landscape design.

To compute the integrated probability of development, as in the urban growth model, we divide the landscape up into non-overlapping square application "panes" (5 km on a side) and then for each pane we define an overlapping "window" of nine panes centered on the focal pane. In contrast to the urban growth model, we apply the process described below to each of the nine overlapping windows (instead of panes) and compute the weighted average (as described below) for the each pane. This overlapping window approach ensures that we avoid ending up with arbitrary edges or abrupt changes in probability of development along the edges of the application panes. While this computationally intensive approach is perhaps ideal for the urban growth model as well, we deemed it not that important in the context of simulating individual disturbance. However, avoiding the arbitrary edges of the application panes for the integrated probability of development surface was deemed important enough to warrant the additional computational cost.

Briefly, we determine the relative probability of each development transition (i.e., suitability surface) based on the landscape condition in 2010 (as in the urban growth model for the first timestep) and adjust these probabilities to reflect the cumulative 70-year (2010-2080) allocation of demand for new development among application windows within each CBSA or county — only here we allocate to the application window instead of pane because of the overlapping window approach described above. A detailed description of the process follows:

1. For each transition type and overlapping application window, we do the following:
    a. As in the urban growth model, determine the relative probability of development (i.e., suitability) for each 30 m cell based on the weighted-average logistic regression model (see Suitability above);
    b. mask out undevelopable cells, including roads, wetlands, conserved land, already developed lands, etc. to enforce a zero probability of development for these cells;
    c. divide the window by the sum of the probabilities in the window. Note, this normalizes the probabilities so that they sum to 1 for the window;
    d. calculate the probability of each cell being developed (P_t) given the number of cells of development for this transition allocated to the window as:

$$P_t = 1 - (1 - P_t^*)^n$$

    where P_t* = the normalized probability from step c for the t^{th} transition type, and n is the cumulative number of cells allocated to the window for this transition for the entire 70-year period (2010-2080). For each window we now have the actual probability of development for this transition occurring sometime between 2010-2080 at the 30 m cell level. Note, here the probability applies to the cell as if we

were to develop cells individually, rather than in patches (as actually happens in the real world and in our urban growth model), which is an approximation that we found acceptable for this particular application;

e. calculate weights for each cell in the window based on a logistic function of the distance to the center of the window as:

$$w_t = \frac{1}{1 + e^{-(b \cdot (x - c))}}$$

where: b=0.05, c=0.8*s, s=166 (size of the pane in cells), and x=distance (m) to the center of the window (**Fig. 3**);

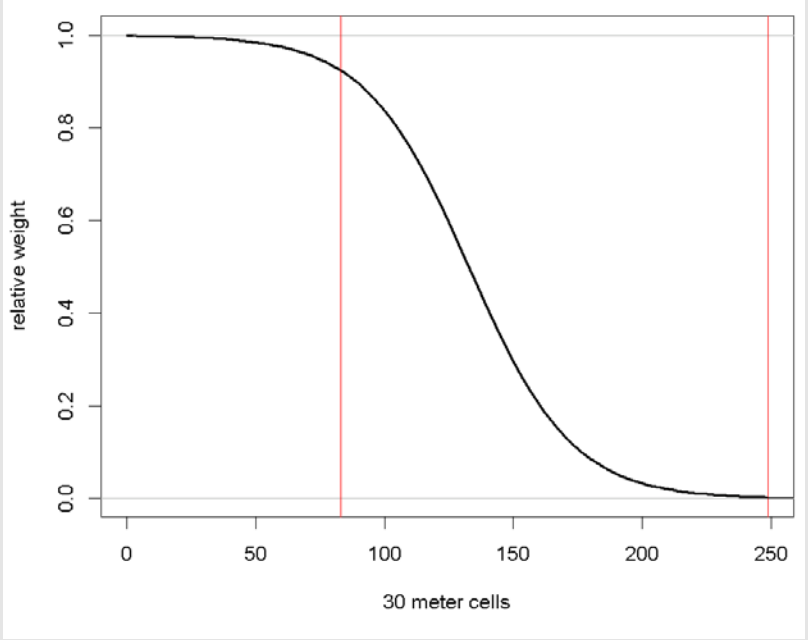

Figure 3. Logistic weighting function used to weight each 30 m cell in an application window (~15 km on a side) to create a weighted average probability of development for each transition type (see text for details).

f. add the weights for the window to a grid of total weights used;

g. add the product of the weights and the probability surface for the window (from step d) to an intermediate grid;

h. repeat steps a-g for every window;

i. divide the intermediate grid (step g) by the weights grid (g) to get a continuous probability of development surface for this transition. Note, in this grid each cell is a weighted average of the probability of development in the nine windows that overlap it; the weights being a logistic function of the distance to the center of each window;

2. repeat step 1 for each of the six transition types; and

3. calculate a weighted joint probability of development (across all transition types) as:

$$1 - \sum_{t=1}^{6} \left(1 - (P_t \cdot W_t)\right)$$

where the weights, W_t, are given as:

- new low=0.5
- new medium=0.8
- new high=1.0

- low to medium=0.3
- low to high=0.5
- medium to high=0.2).

GIS metadata

This data product is distributed as a geotiff raster (30 m cells) where the cell value equals the integrated probability of development (0-1). This product can be found at McGarigal et al (2017).

Literature Cited

McGarigal K, Compton BW, Plunkett EB, DeLuca WV, and Grand J. 2017. Designing sustainable landscapes products, including technical documentation and data products. https://scholarworks.umass.edu/designing_sustainable_landscapes/

Wear DN. 2011. Forecasts of county-level land uses under three future scenarios: a technical document supporting the Forest Service 2010 RPA Assessment. Gen. Tech. Rep. SRS-141. Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station. 41 p.