

Completion of the 2001 National Land Cover Database for the Conterminous United States

Collin Homer, Jon Dewitz, Joyce Fry, Michael Coan,
Nazmul Hossain, Charles Larson, Nate Herold,
Alexa McKerrow, J. Nick VanDriel, and James Wickham

Introduction

Appropriate and relevant land cover information is increasingly required by a broad spectrum of scientific, economic and governmental applications as essential input to analyze such issues as assessing ecosystem status and health, understanding spatial patterns of biodiversity and developing land management policy. The publication of the first National Land Cover Dataset (NLCD 1992) (Vogelmann *et al.* 2001) created a 30-meter resolution land cover data layer over the conterminous United States from circa 1992 Landsat Thematic Mapper (TM) imagery. Information from this original NLCD 1992 has been used in thousands of applications in the private, public, and academic sectors ranging from assisting in placing cell phone towers to tracking how diseases spread. The national consistency of this information has provided critical analysis for many national applications such as the Heinz Center's *State of the Nation's Ecosystems* (Heinz Center 2002), the Environmental Protection Agency's *Draft Report on the Environment* (USEPA 2003) and the U.S. Geological Survey National Water Quality Assessment program.

Starting in 1999, new research was undertaken to expand and update NLCD 1992 into a full scale land cover database (with multiple instead of single products), and to produce it across all 50 states and Puerto Rico (Homer *et al.* 2004). This new database is called the National Land Cover Database 2001 (the 2001 refers to the nominal year from which most of the Landsat 5 and Landsat 7 imagery was acquired) and has been under production for 6 years. This article announces the completion of NLCD 2001 for the conterminous United States, with products that can identify one of 16 classes of land cover, the percent tree canopy, and the percent urban imperviousness for every 30-meter cell in the conterminous 48 states (*approximately 27 billion cells*).

Both NLCD 1992 and NLCD 2001 have been produced and funded through an umbrella organization called the Multi-Resolution Land Characteristics Consortium (MRLC). This Consortium consists of 13 programs across 10 Federal agencies that require

landcover data for addressing their agency needs (www.mrlc.gov). MRLC provided the umbrella to coordinate multi-agency NLCD mapping production and funding contributions. In addition to NLCD data, MRLC also offers approximately 6,200 terrain corrected Landsat 5 TM and Landsat 7 Enhanced Thematic Mapper (ETM+) scenes spanning growing season dates from 1982-2006 which are available for public web-enabled download from www.mrlc.gov. MRLC represents an excellent example of Federal government collaboration across many agencies to synergistically develop important geo-spatial data for the Nation.

Methods

NLCD 2001 was generated according to method protocols outlined in Homer *et al.* (2004) using 65 mapping zones for the conterminous United States. Production occurred across 12 mapping teams from both the government and private sector. In order to ensure product consistency among teams, products were generated using a standardized process spanning data preparation, classification and quality control. USGS EROS was responsible for oversight of product development, data preparation, classification training, quality control and product synthesis. A brief overview of standard methods is presented in the following section; slight variations for unique issues in individual zones are not represented.

Source Data Preparation

All NLCD 2001 products were generated from a standardized set of data layers mosaiced by mapping zone. Typical zonal layers included multi-season Landsat 5 and Landsat 7 imagery centered on a nominal collection year of 2001, and Digital Elevation Model based derivatives (Figure 1). This standard set of mosaiced zonal layer stacks often consisted of 18 or more layers that provided the best available data resources to derive the desired products. Application of layers could vary slightly, as mapping protocols were

continued on page 338

shifted to meet unique regional conditions. All data were geo-registered to the Albers equal area projection grid, and resampled to 30m grid cells.

Land Cover Classification

The landcover classification was accomplished using commercial decision tree (DT) software called See5* (Quinlan 1993) applied to zonal layer stacks prepared for each mapping zone (Figure 1). In addition to the See5 software, an interface was created for ERDAS IMAGINE* to extrapolate derived DT models into classified pixels. DT is a supervised classification method that relies on large amounts of training data, which was initially collected from a variety of sources including high-resolution orthoimagery, local datasets, field collected points, and Forest Inventory Analysis (FIA) plot data. In many mapping zones, training data collection took advantage of existing regional land cover maps (e.g. NLCD 1992, Gap Analysis Program (GAP), and National Agricultural Statistics Service (NASS) cropland data) to improve classification efficiency. Predictions from these multiple products were compared to find areas of agreement, then spatially sampled to avoid transitional or edge pixels to generate training samples randomly across classes in proportion to their population. The result produced evenly distributed training data allowing for optimization of the DT models. Training data were used to map all land cover classes except for the four urban classes which were derived from thresholding of the imperviousness data product (see Table 1).

Once an initial classification was completed, typically a number of subsequent DT iterations were necessary to improve the classification result. A series of scripts specifically written for this project were employed to gauge success and make adjustments to the See5 data file as required to generate an acceptable map. Once the product had evolved as far as DT methods could take it, additional localized modeling and hand-editing were typically required to produce the final product. However, localized modeling and hand-editing affected proportionately few pixels.

Imperviousness and Canopy Classifications

Imperviousness and tree canopy were classified using commercial regression tree (RT) software called Cubist* (Yang *et al.* 2002). Training data were generally derived from 1-m resolution Digital Orthoimagery Quarter Quadrangles (DOQQs), classified categorically into canopy/non-canopy, or impervious/non-impervious for

each 1-m pixel, and subsequently resampled to 30-m grid proportions. Early products were classified with one or two DOQQs of 6-8 square kilometers per Landsat path/row, but in later products accuracy and prediction quality was improved by distributing 3-4 smaller chips of 1-4 square kilometers in size throughout the Landsat scene. Increasing the sampling frequency not only improved the reliability of training distributions to capture the total range of zonal estimates, but also improved classification efficiency and reduced costs. The completed training images were then extrapolated across mapping zones using RT models, to derive continuous canopy and imperviousness estimates.

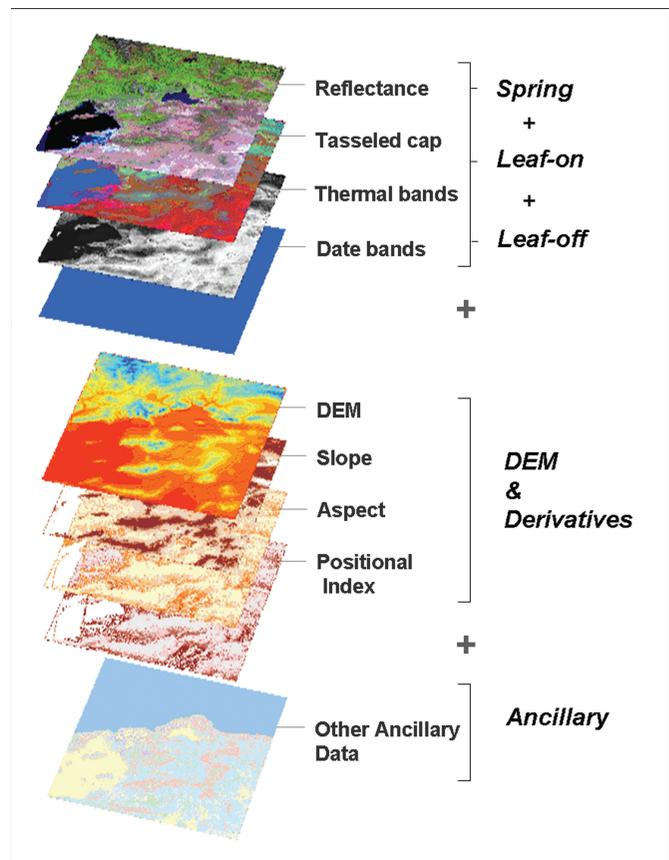


Figure 1. Mapping Zone Input Layers used to Derive NLCD 2001 Products.

Table 1. Tree Canopy and Urban Impervious CONUS Predictions, by 10 % Groupings.

Percent Group	Canopy		Impervious		NLCD Urban Class
	Area In Sq. km	Percentage	Area In Sq. km	Percentage	
1-10	52873.56	1.96	215394.65	47.13	Developed, Open Space (21)
11-20	104252.83	3.87	75465.8	16.51	
21-30	143369.52	5.33	48774.54	10.67	Developed, Low Intensity (22)
31-40	161737.93	6.01	34976.8	7.65	
41-50	179218.63	6.66	25995.62	5.69	
51-60	199291.24	7.4	18995.29	4.16	Developed, Medium Intensity (23)
61-70	240062.08	8.92	13317.38	2.91	
71-80	392854.96	14.59	9579.07	2.1	
81-90	777914.63	28.9	7750.25	1.7	Developed, High Intensity (24)
91-100	440412.98	16.36	6809.66	1.49	
Total	2691988.36	100	457059.06	100	

Estimates for canopy or imperviousness were created on all pixels, with a subsequent masking strategy employed to reduce errors of commission on estimate areas with spectrally similar features difficult to discriminate accurately (e.g., shrub and grass areas for canopy, and bare agriculture fields for imperviousness). Canopy masks were generated using See5, spectral classifications, and some image segmentation-based classifications. These masks were later harmonized with the completed land cover, by adding any forested land cover not included in the original canopy mask. Imperviousness masks were produced by combining GIS layers of road density buffers, city lights, spectral classifications, and image segmentation-based classifications. Completed masks were hand-edited to ensure an accurate inclusion of urban areas. NLCD 2001 canopy and imperviousness products are masked versions of the predictions, unmasked versions are available only by special request.

Post-processing

When landcover modeling was completed, the final product was aggregated to a one acre minimum mapping unit (five TM pixels)

using a “smart eliminate” aggregation algorithm. This algorithm uses eight-corner connectivity from a central pixel to allow non-linear features like roads and streams to remain intact, and accesses a weighting table to allow “smart” decisions on a dissolve protocol. In a few landcover zones, higher minimum mapping unit thresholds were applied to agricultural classes to reduce commission errors in problematic areas. Canopy and imperviousness products underwent no aggregation.

While every effort was made to maintain consistency in classification between mapping zones during production, some edge matching was required to merge the 65 mapping zones. Because each mapping zone was initially produced with a three kilometer boundary buffer, a six kilometer overlap was available between mapping zones for edge-matching. During the edge-matching process, each zone boundary was scrutinized and adjusted to minimize any classification inconsistencies in the final seamless product.

continued on page 340

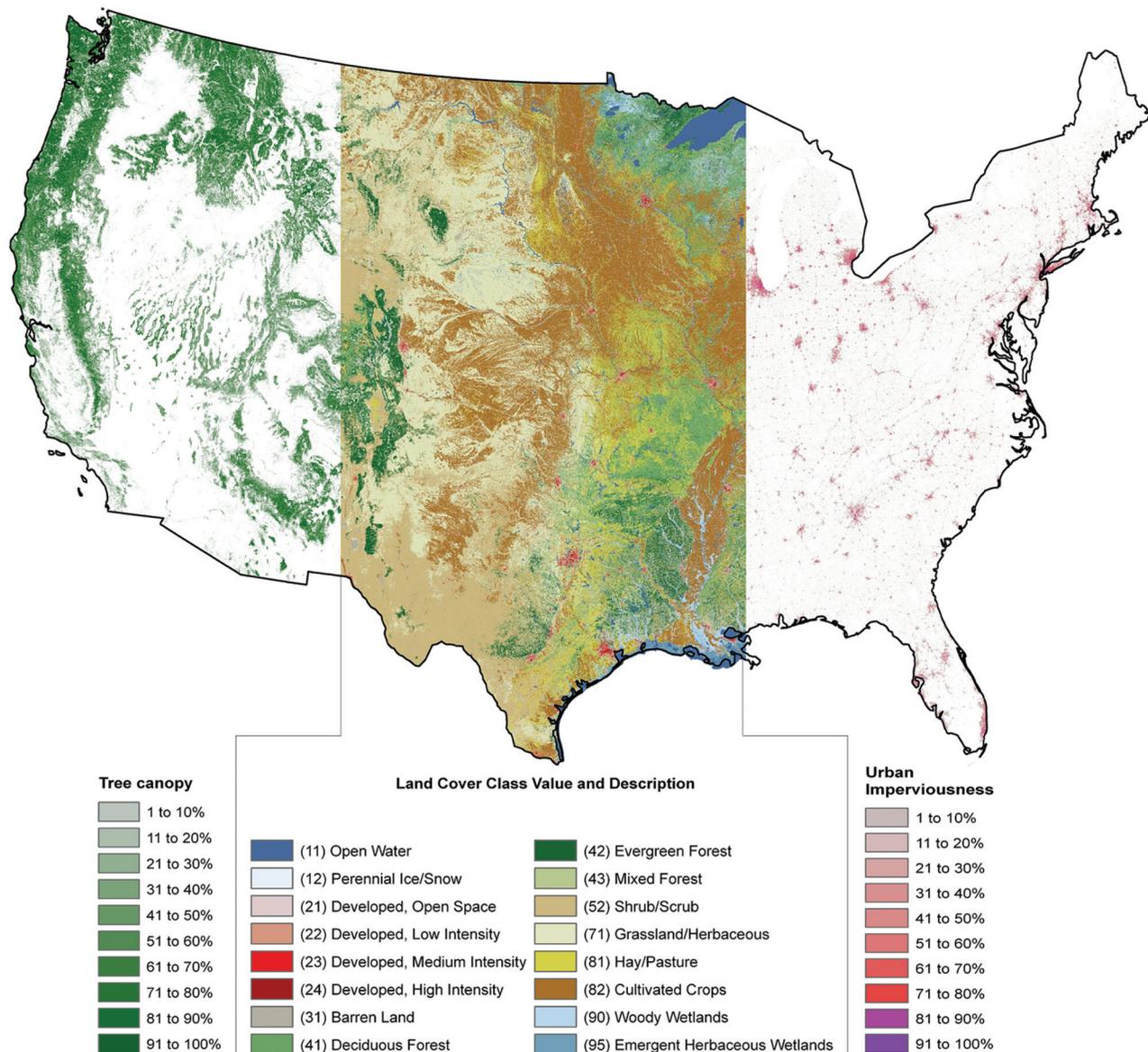


Figure 2. NLCD 2001 Tree Canopy, Land Cover and Urban Imperviousness Products.

Results

Land Cover

Sixteen classes of land cover were modeled over the conterminous United States at a 30m cell size with a 1 acre minimum mapping unit (Figure 2). Proportionately, the rarest class was perennial ice/snow at 0.02% of the total area and shrub/scrub the most common class at 21.03% of the total area (Figure 3). Full legend class descriptions were published in Homer *et al.* 2004.

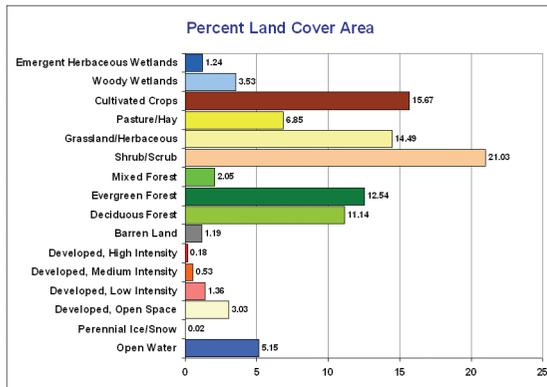


Figure 3. NLCD 2001 Land Cover Proportions, by Class.

Imperviousness and Canopy Classification

Continuous predictions from 1-100% were produced for tree canopy and urban imperviousness over the conterminous United States at 30m cell resolution (Figure 2). As displayed in Table 1, canopy groupings from 91-100% represent the largest proportion at 16.36% of total area, and canopy groupings from 1-10% represent the smallest proportion at 1.96% of total area. Imperviousness groupings from 1-10% represent the largest proportion at 47.13% of total area (likely due to the large number of tertiary roads outside of urban areas), and imperviousness groupings from 91-100% represent the smallest proportion at 1.49% of total area. These products were produced using a 1% interval to distinguish among classes, the examples in Figure 2 and results in Table 1 are summarized in larger intervals for reporting convenience.

Product Accuracy

No formal accuracy assessment of the NLCD 2001 products has yet been completed; however, one is planned in the near future based on the design outlined in Stehman *et al.* (unpublished). In the meantime, users can gain initial feedback on product accuracy from the cross-validation estimate of product accuracy provided from the DT/RT algorithms employed in NLCD 2001 modeling. Typically, a 10-fold cross-validation was conducted by dividing the entire training data set into 10 subsets of equal size. For each model run, an accuracy estimate was derived using one subset to evaluate the model prediction derived from the other 9, with the process repeated 10 times. After all 10 runs, the error estimate is computed and represents the reported number. While cross-validation can potentially provide relatively reliable estimates of model prediction accuracy if reference data follow a probability sampling design, we make no guarantee this criteria was met. Users are cautioned that these cross-validation results represent only first-order estimates of data quality, and should not be considered a formal accuracy assessment. Some results may report optimistic accuracies eventually unsubstantiated by a formal assessment.

Cross validation accuracy of the land cover product was weighted by class occurrence in each mapping zone. Accuracy estimates

across mapping zones ranged from 70% to 98%, with an overall average accuracy across all mapping zones of 83.9%. Because of the continuous estimate format for canopy and imperviousness, accuracy is reported as an average error estimate. Average error is derived from a zonal extrapolation, and allows the user to estimate the magnitude a cell value may deviate from the prediction. Canopy product average errors range from 6% to 17% deviation from prediction, while imperviousness average error ranged from 4% to 17%. Mapping zone-dependant cross validation error estimates are reported in the NLCD 2001 metadata.

Discussion

The major value of NLCD 2001 lies in its ability to provide a complete consistent coverage of the nation's land cover on relevant imagery, and serve as a resource for regional to national scale applications. Users should be cautioned this type of database is designed to meet regional and national requirements, and is not designed for local application (e.g., county level use). We recognize many users will desire to modify or develop value-added steps to customize NLCD 2001 data products for their specific applications. This philosophy was accommodated in the original design (Homer *et al.* 2004), with results including; intermediate data layers including image transformations, ancillary information and multi-season mosaics organized by mapping zone (Figure 1), three products of land cover, tree canopy and urban imperviousness designed to each offer unique information and be used either independently or synergistically (Figure 2), and comprehensive metadata provided to document data production procedures. The development of these database products resulted in the successful delivery of baseline imagery, products and modeling information to help optimize other national programs and secure their cooperation. Additional land cover products of classification rules and classification confidence data sets were originally envisioned to provide additional information about the land cover classification. Production of those two additional data sets has not yet been funded.

Comparison of NLCD 1992 and NLCD 2001

New improvements in mapping methodology, input data, and minor mapping legend modification confound comparison between NLCD 1992 and NLCD 2001, and direct comparison of these two independently created land cover products is not recommended. Users are likely to discover that differences in the methodology used to produce the two products, overwhelm true differences due to land cover change. However, the NLCD design team has developed a "bridge product" to aid land cover change analysis between the two eras. Because early research showed that a comparison at the full NLCD legend resolution would cause unacceptable error, this product is derived at the more achievable Anderson Level I scale across 8 broad land cover classes (Coan *et al.*, in prep.).

This product recreates both a 1992 and 2001 Anderson Level I classification using processing tools developed from NLCD 2001. A multi-stage processing method utilizes areas of agreement between NLCD 1992 and NLCD 2001 land cover to derive training, then classifies both eras of land cover with a DT classifier, and finally filters subsequent products with DT confidence parameters to identify changed pixels. The pixels subsequently go through a labeling process to identify the "from-to" change classification code. Initial accuracy analysis of this method showed an overall 94.2% agreement on change/no-change pixel identification against change/no-change pixels identified using traditional methods. This product is scheduled to be completed for the conterminous United States by the end of 2007.

Conclusions

The completion of NLCD 2001 for the conterminous United States has been widely anticipated by users seeking updated land cover information and provides a modernized version of our Nation's land cover from the original NLCD 1992. The NLCD 2001 product set and mapping tools are available via Web-enabled file download from the MRLC Consortium website (www.mrlc.gov) with options for both Dynamic Download (user-defined download areas) and FTP Download by zonal groupings. The files are available in GeoTIFF, ArcGRID, or BIL formats in most cases; however, the FTP Zonal groupings are limited to ERDAS format. Mapping zone-level metadata is supplied with all downloads, with standard formats including HTML, XML and TXT. All NLCD 2001 product sets are distributed at 30-meter resolution in the NLCD-standard NAD 83, Albers equal area conic projection.

NLCD 2001 data for Alaska, Hawaii, and Puerto Rico will be completed by December of 2007, which will then represent the first compilation of nationwide land cover ever produced at 30-meter resolution. Future updates of NLCD 2001 are planned to continue support of land cover requirements across the Nation, possibly at an increased temporal frequency.

References

- Coan, M., Fry, J., Homer, C., Larson, C., 2007. (Manuscript in Preparation). Developing a Land Cover Change Product from Two Generations of the National Land Cover Database (1992-2001)
- Heinz Center (The H. John Heinz III Center for Science, Economics, and the Environment). 2002. *The State of the Nation's Ecosystems: Measuring the Lands, Waters, and Living Resources of the United States*. Cambridge University Press, Cambridge, UK.
- Homer, C. C. Huang, L. Yang, B. Wylie and M. Coan. 2004. Development of a 2001 National Landcover Database for the United States. *Photogrammetric Engineering and Remote Sensing*, Vol. 70, No. 7, pp 829-840
- Quinlan, J. R., 1993. C4.5: *Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California.
- Stehman, S.V., J.D. Wickham, T.G. Wade, J.H. Smith. (unpubl.) (accepted, pending revision). Designing a multi-object, multi-support accuracy assessment of the 2001 National Land Cover Data (NLCD 2001) of the United States. *Photogrammetric Engineering and Remote Sensing*.
- USEPA (U.S.Environmental Protection Agency). 2003. Draft Report on the Environment 2003. <http://epa.gov/indicators/roe/html/roePDF.htm>.
- Vogelmann, J.E., S.M. Howard, L. Yang, C. R. Larson, B. K. Wylie, and J. N. Van Driel, 2001, Completion of the 1990's National Land

Cover Data Set for the conterminous United States, *Photogrammetric Engineering and Remote Sensing* 67:650-662.

Yang, L, C. Huang, C. Homer, B. Wylie, and M. Coan., 2002. An approach for mapping large-area impervious surfaces: Synergistic use of Landsat 7 ETM+ and high spatial resolution imagery, *Canadian Journal of Remote Sensing* 29(2):230-240.

Acknowledgements

This research reflects an enormous team effort across many organizations and individuals. Because of the number of individuals involved, they cannot be properly acknowledged here. However, we do acknowledge the many organizations that made this work possible. We especially acknowledge the support of the individuals and agencies of the MRLC Consortium, the many NLCD 2001 Federal mapping teams from the USGS, NOAA, and USFS, and the private sector mapping teams of MDA Federal and Sanborn who performed work under several USGS and NOAA contracts. This study is made possible in part by the SAIC Corporation under U.S. Geological Survey contract 03CRCN0001. Additional thanks is given to all of the organizations, agencies and individuals who provided training data for NLCD 2001 modeling – especially the contribution of the USDA-FS FIA data to develop forest classifications across many zones.

Endnote

The use of any trade, product or firm name is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Authors

Collin Homer, J. Nick VanDriel

U.S. Geological Survey (USGS) Center for Earth Resources Observation and Science (EROS), Sioux Falls, SD 57198

Jon Dewitz, Joyce Fry, Michael Coan, Nazmul Hossain, and Charles Larson

SAIC, contractor to USGS EROS, Sioux Falls, SD 57198

Nate Herold

NOAA Coastal Services Center, 2234 South Hobson Avenue., Charleston, SC 29405

Alexa McKerrow

Southeast Gap Analysis Project, Department of Zoology, North Carolina State University, Raleigh, NC 27695-7617

James Wickham

U.S. Environmental Protection Agency (EPA), National Exposure Research Laboratory (E243-05), Research Triangle Park, NC 27711