City-City Correlations to Introduce Galaxy-Galaxy Correlations

Daniel M. Smith, Jr.

South Carolina State University

Department of Biological and Physical Sciences

Orangeburg, SC 29117

(Dated: May 10, 2012)

Abstract

Large Scale Structure of the universe, a consequence of the Big Bang model and its refinements, is commonly characterized by the two-point galaxy-galaxy correlation function. But the meaning of the correlation function is somewhat abstract because it does not have a ready physical analog. To lower the barrier to comprehension, this work computes the 2-dimensional, two-point city-city correlation function for three populous regions of the United States. The favorable correspondence between this function and more direct measures of city radii and city-city distances helps to clarify the meaning of the correlation function in 3-dimensions for galaxies. A rudimentary analysis of Sloan Digital Sky Survey data demonstrates that the correlation function graphs for galaxies and cities are quite similar and that at small enough distances a power law can represent them both. Part of this work has been adapted for a lab suitable for undergraduates.

I. INTRODUCTION

Dark energy (Λ), cold dark matter (CDM), photons, and baryons all determine the expansion of the universe within the context of the Big Bang model (called the ΛCDM model),[1] and they also determine the resulting Large Scale Structure (LSS) of the universe as revealed by galaxies. An important tool in characterizing the LSS is the galaxy-galaxy two-point correlation function[2] whose applications include determining the dark matter percentage of the universe,[3,4] and demonstrating galaxy clustering due to an acoustic wave—protons coupled to photons streaming out of dark matter potential wells.[5] Although galaxy data are available on-line, convenient for download and analysis, the concept of the correlation function might seem obscure to the uninitiated, thus motivating the current work to calculate, for three groups of United States cities, the 2-dimensional, two-point correlation function that can be easily interpreted because of readily available city data. Then non-experts can more readily interpret the 3-dimensional two-point galaxy-galaxy correlation function calculated from Sloan Digital Sky Survey (SDSS) data or from other surveys. The question posed and answered in this work is the following: in what sense, mathematically, does the clustering of galaxies in Figure I resemble the clustering of cities in Figure 2?

The two-point correlation function is briefly introduced in Section II of this work, and the method of estimating error is explained in Section III. The correlation function calculation for cities is described in Section IV, and Section V contains a rudimentary correlation function calculation for galaxies. An adaptation of this work for an undergraduate lab is discussed in Section VI. Section VII summarizes the work.

II. TWO-POINT CORRELATION FUNCTION

The probability of finding a galaxy in each of the randomly placed volume elements dV_1 and dV_2 in a uniform distribution of galaxies is

$$dP_1 dP_2 = \frac{n_1 n_2}{N^2} dV_1 dV_2, \tag{1}$$

where n_1 and n_2 are galaxy number densities, and N is the number of galaxies in the volume considered. In the presence of gravity, the probability of finding a galaxy in each of the volume elements dV_1 and dV_2 separated by a distance s is

$$dP_1 dP_2 = \frac{n_1 n_2}{N^2}(1 + \xi(s))dV_1 dV_2, \tag{2}$$

where $\xi(s)$ is the two-point correlation function, representing the enhancement of galaxy clustering due to gravity. This work uses the Davis-Peebles estimator of the correlation function:[6,7]

$$\xi(s) = \frac{2N_R}{N_D - 1}\frac{\langle DD(s)\rangle}{\langle DR(s)\rangle} - 1, \tag{3}$$

where N_R and N_D are the number of galaxies in the data and random catalogs respectively, $\langle DD(s)\rangle$= number of galaxy-galaxy pairs at separation s, and $\langle DR(s)\rangle$= number of random-galaxy pairs at separation s for a random distribution of galaxies. Although estimators for $\xi(s)$ have been defined in a few other ways,[7–9] the Davis-Peebles estimator is adopted for this work because for cities it is more consistent with directly measured values for city-city distances and city radii under the condition $N_D/N_R = 1$ used throughout this work. Random catalogs generated, explained in more detail below, are constrained by the same geometry as the actual data. Equations (1)–(3) are replaced by their 2-dimensional analogues in the analysis of cities.

FIG. 1. Plot of 9,659 galaxies from the SDSS Data Release 7 for $0 < z < 0.05$ and $-3° < dec < 3°$.

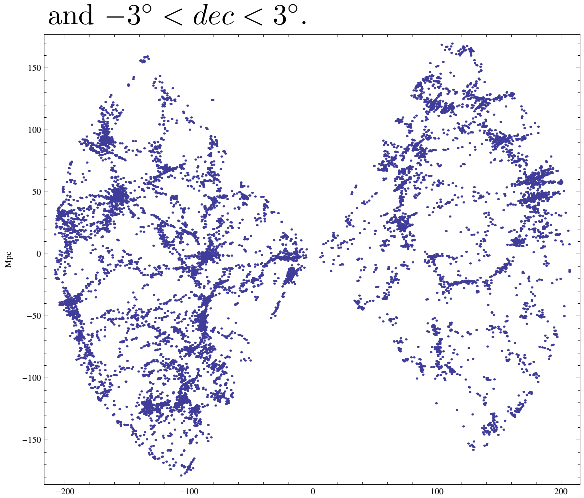
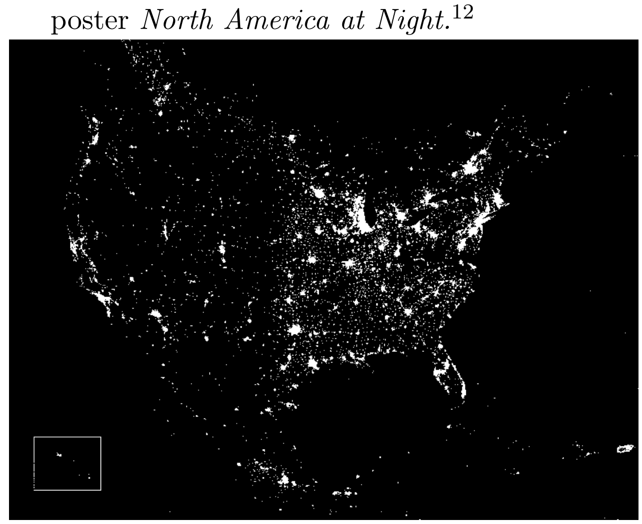


FIG. 2. Clustering of Midwest, Southwest, and Southern cities is analyzed from the poster *North America at Night*.[12]



III. ERROR ESTIMATE

The so-called jackknife method is used to estimate the standard deviation in the correlation function calculation.[10,11] The advantage of this method of error estimation is that it entails no assumption that the data have a Gaussian (or any other) distribution about some mean. The data and random values are divided up into N equal parts and the correlation function is calculated N times, each time leaving out one part of the data and the associated random values. Any one of these correlation functions is designated $\xi_i(s)$, and the average of all N of them is $\bar{\xi}(s)$. Then the standard deviation, σ, is given by:[10]

$$\sigma = \sqrt{\frac{N-1}{N}\sum_{i=1}^{N}\left(\bar{\xi}(s) - \xi_i(s)\right)^2}\,. \tag{4}$$

IV. CITY-CITY TWO-POINT CORRELATION FUNCTION

The formalism of the previous two sections was adapted to the problem of determining the city-city two-point correlation function for three groups of cities. The composite satellite image poster (Fig. 2), *North America at Night*,[12] was digitized, and cities represented by a blob of light were identified by comparison with a standard geographical map. Three regions were chosen as representative of a cluster of cities. The criteria for choosing a region were (1) that its shape must be square (for ease of analysis), and (2) that two other non-overlapping regions could be chosen of the same size and approximate light density. The Midwest, Southwest, and Southern regions of the United States chosen and described in more detail below were approximately 160,000 miles^2 in area.

To perform a correlation function analysis comparable to that for galaxies, the square image was converted via software to a matrix of 1s and 0s, with one white pixel on the night map considered analogous to one galaxy. An algorithm counted the number of 1s so that the same number of white pixels could be generated in a square random-light map for each region, required for equation (3). Each city cluster was divided into eight regions of equal area for the jackknife error analysis, then the correlation function was calculated eight times by successively omitting one region at a time then averaging the eight results. To enable interpretation of the correlation function, pixel positions of the center and edge of each city were recorded directly from the computer screen by using the *Mathematica* Drawing

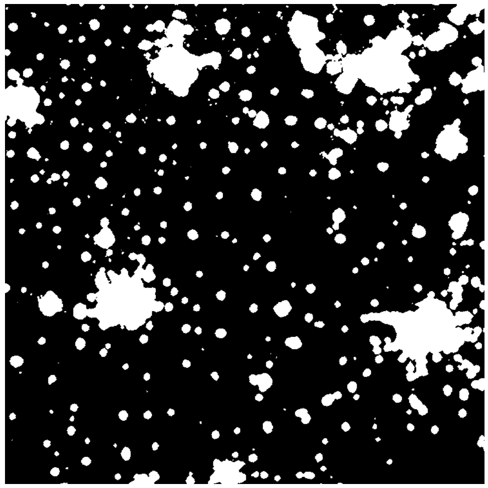


FIG. 3. The correlation function is calculated for this group of Midwest cities. Clockwise, from the upper left corner: Omaha, Des Moines, Cedar Rapids, Davenport, St. Louis, Springfield, and Kansas City.

Tools,[13] enabling calculations of city radii and city-city distances in pixels. Actual distances (in miles) were found by referencing the latitudes (θ_1 and θ_2) and longitudes (ϕ_1 and ϕ_1) of any two cities (from the CityData function in *Mathematica*) then computing the great circle distance, d, between them:

$$d = (3963\,\text{miles}) \arccos[\sin(\theta_1)\sin(\theta_2) + \cos(\theta_1)\cos(\theta_2)\cos(\phi_1 - \phi_2)], \tag{5}$$

where the value of the earth's equitorial radius is from IERS.[14] Thereby, the conversion factor between pixels and miles was determined to be 0.5 miles/pixel.

Using equation (3), the correlation function is calculated for the Midwest cities shown in Figure 3: Omaha, Nebraska; Des Moines, Cedar Rapids, and Davenport, Iowa; and St. Louis, Springfield and Kansas City, Missouri. The result for 93,216 white pixels is displayed in Figures 4 and 5. In Figure 4, the correlation function, initially positive for short distances from a city's center, eventually goes negative then becomes positive again, reaching its largest value ($\xi = 0.12$) at 242 ± 6 miles. This value compares favorably to the 228 ± 82 miles determined by simply averaging all of the possible distances between all of the cities. In Figure 5, the small s data are fitted to a power law: $\xi(s) = (s/20 \pm 10\ \text{mi})^{-0.6 \pm 0.2}$. The characteristic distance of 20 ± 10 miles is comparable to the average radius (20 ± 5 miles) of all of these Midwest cities as determined by the pixel measurements with conversion factor, as described above.

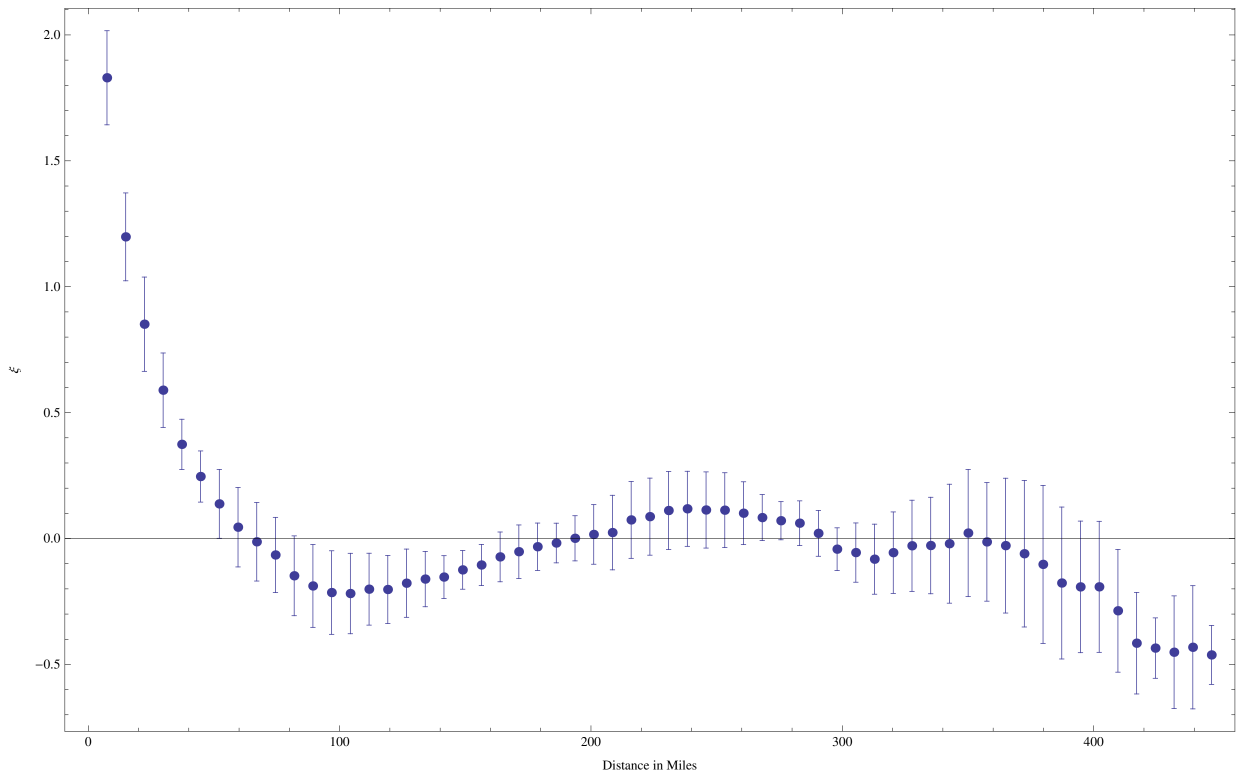


FIG. 4. Two-point correlation function for Midwest cities in Figure 3.

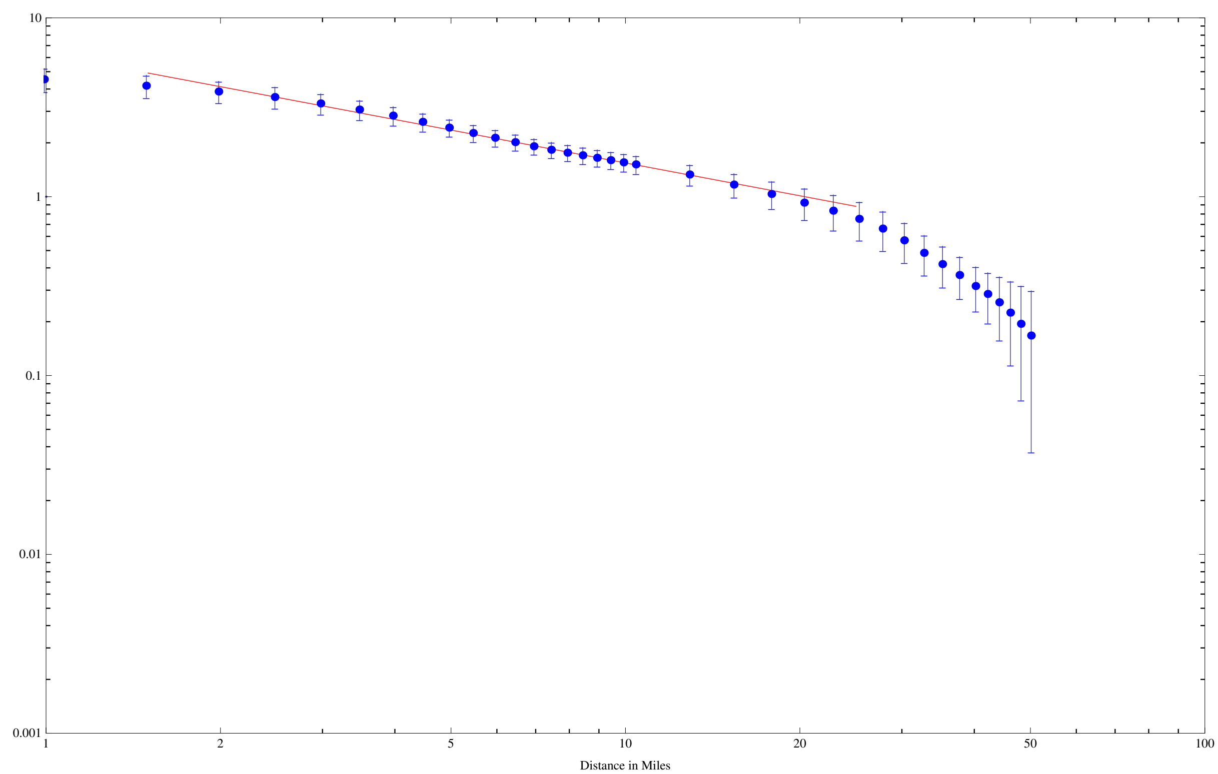


FIG. 5. Power-law fit for Midwest cities correlation function: $\xi(s) = (s/20 \pm 10\text{ mi})^{-0.6\pm0.2}$.

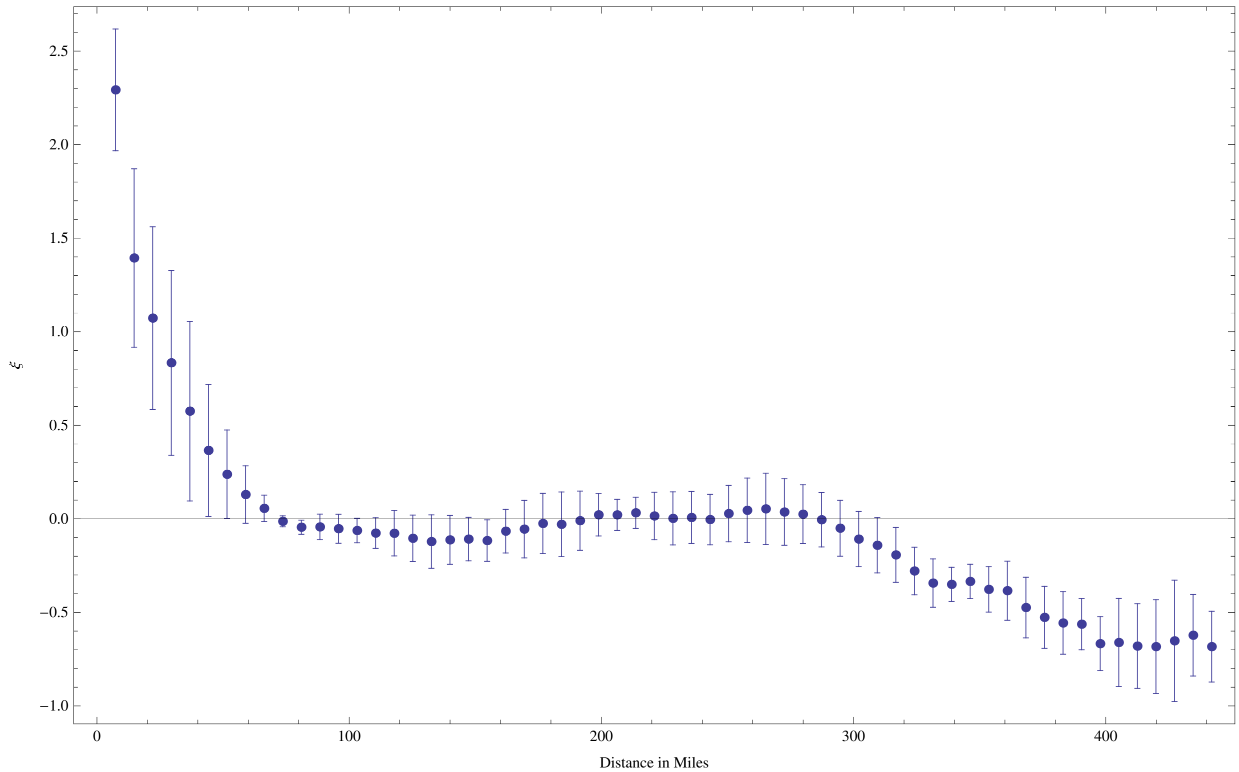


FIG. 6. Two-point correlation function for Southwest cities.

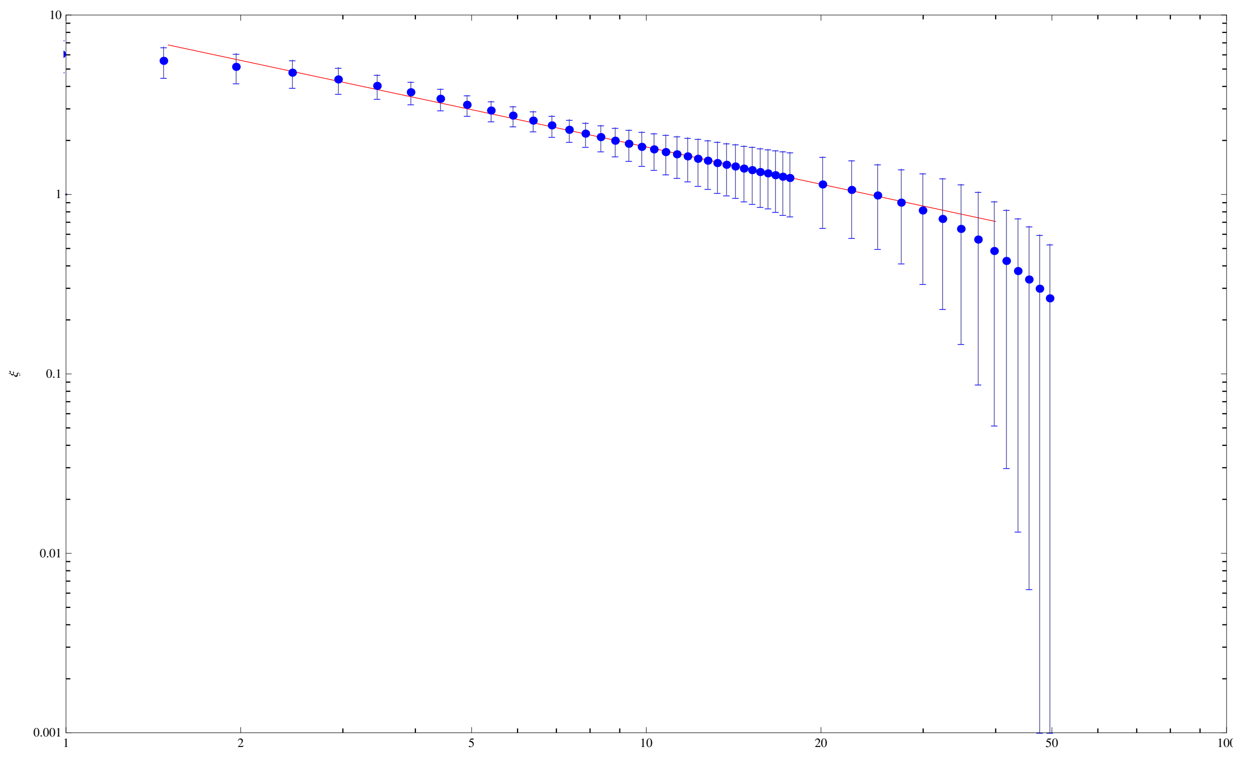


FIG. 7. Power-law fit for Southwest cities correlation function: $\xi(s) = (s/24^{+39}_{-12}\ \mathrm{mi})^{-0.7\pm0.3}$.

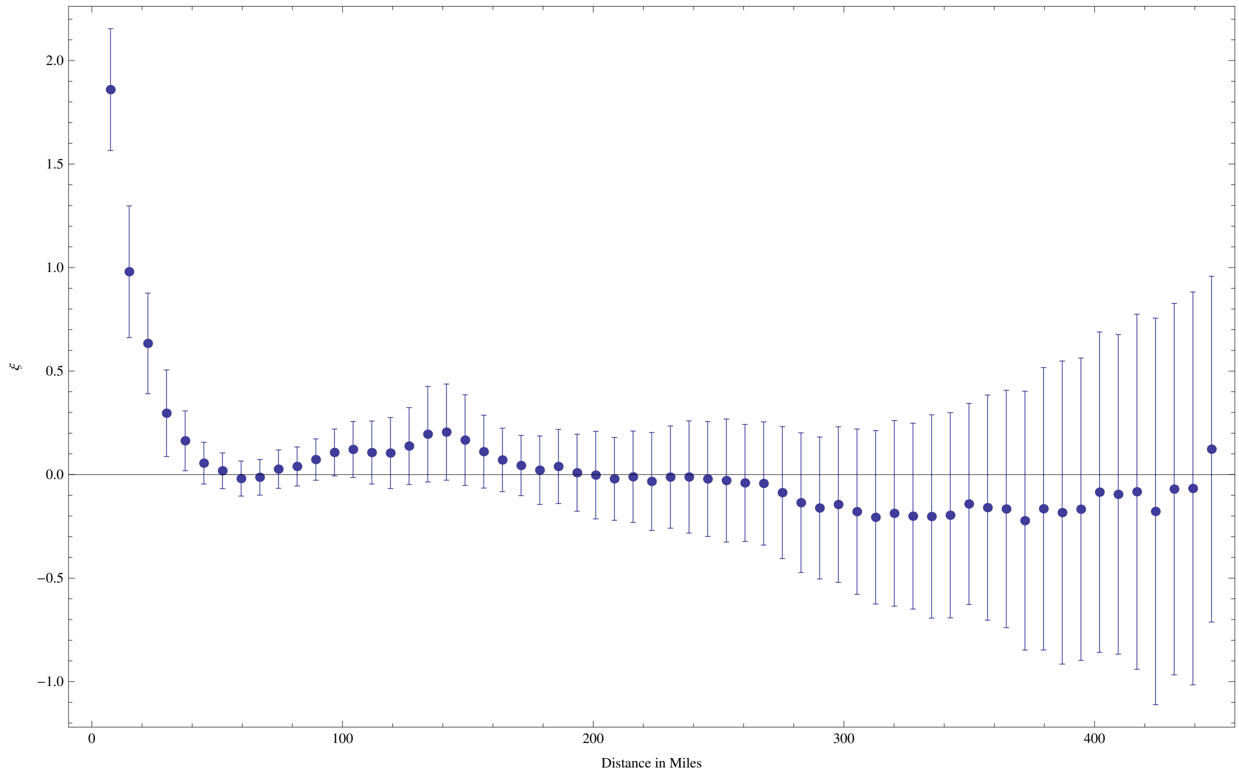


FIG. 8. Two-point correlation function for Southern cities.

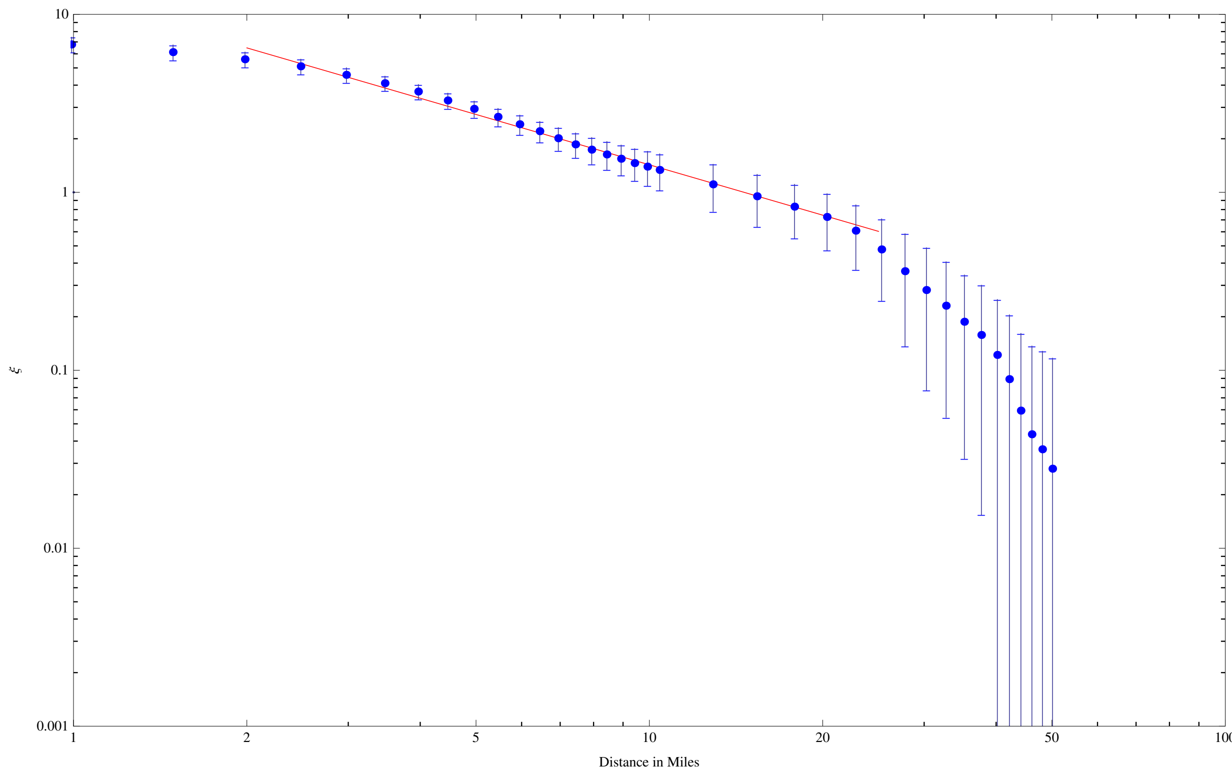


FIG. 9. Power-law fit for Southern cities correlation function: $\xi(s) = (s/15\pm 5 \text{ mi})^{-0.9\pm0.2}$.

A similar analysis, with 72,079 white pixels, is performed for a group of Southwest cities: Oklahoma City, and Tulsa, Oklahoma; Fayetteville and Little Rock, Arkansas; Dallas, Texas; and Shreveport, Louisiana. In Figure 6, the maximum value of ξ away from the city center is $\xi = 0.056$ at 263 ± 6 miles. (Note, however, that there is a relative maximum, $\xi = 0.039$, at 220 miles.) Averaging the distances between all of the cities yields a comparable value of 215 ± 68 miles. The fit to the small s data plotted in Figure 7 is $\xi(s) = (s/24^{+39}_{-12}\ \text{mi})^{-0.7\pm0.3}$. The 24 mile characteristic distance compares favorably with the average of the pixel radii measurements, 22 ± 8 miles.

The correlation function is also calculated for a group of Southern cities, with 66,313 white pixels: Nashville, Knoxville, and Chattanooga, Tennessee; Birmingham and Montgomery, Alabama; and Atlanta, Georgia. In the Figure 8 plot, the maximum value of $\xi = 0.16$ at 149 ± 4 miles is comparable to the average distance between these cities of 155 ± 58 miles. The fit to the small s data plotted in Figure 9 is $\xi(s) = (s/15 \pm 5\text{mi})^{-0.9\pm0.2}$, where the characteristic distance of 15 ± 5 miles is comparable to the average measured city radius of 14 ± 8 miles.

Average distances between cities, and average city radii for all three groups of cities are summarized in Table I along with the parameters for a power law fit. Note that in all cases, the characteristic distances in the power law correspond to the average measured city radii, but Figures 4, 6, and 8 show that at that distance $\xi \simeq 0.9$. Depending upon the city group, $\xi = 0$ in the 40–80 mile range. This indicates that the characteristic distances account for the central city lights only, and that the correlation remains positive beyond these distances due to suburban lights.

V. GALAXY-GALAXY TWO-POINT CORRELATION FUNCTION

The procedure for an approximate calculation of the galaxy-galaxy two-point correlation function follows that of the previous section and the SDSS Early Data Release (EDR) work of Zehavi *et al.*[15] SDSS Data Release 7 (DR7)[16] is analyzed for galaxies with redshift $0.019 \leq z \leq 0.13$ in the 623.0 nm r band (r^* in Ref. 15) for apparent magnitudes $14.5 \leq r \leq 17.6$. These galaxies are subject to the further restriction $-22 < M_r < -19$ in absolute magnitude,[15,17,18] resulting in a data set of 330,393 galaxies. For the jackknife analysis of Section III, the data set is divided into ten contiguous redshift bins, with an equal number of galaxies in each bin. A catalog random in z (within the above limits) is also generated, matching the bin geometry and the number of data values. Then equation (3) is used to calculate the correlation function by successively excluding galaxies and random values one bin at a time, after which the ten correlation functions are averaged.

The aim here is to do a rough calculation of the correlation function, so to save computation time, random catalogs in this work contain the same number of galaxies as the real catalog, unlike the thorough analysis of the Ref. 15 workers who generate random catalogs with ten times the number of actual galaxies. Fiber collisions (due to two galaxy images closer than 55″), and angular and radial selection functions require additional corrections, described in Ref. 15, that were omitted here. Although the correlation function calculated without these corrections is in fair agreement with the earlier work, the standard deviation from equation (4) is comparatively large as noted below. But the calculation here is adequate for our purpose of demonstrating the similarity of galaxy clustering to city clustering.

Figure 10 illustrates that the correlation function calculated in this work is in fair agreement with the thorough analysis of Ref. 15. But the large standard deviation for ξ shown in Figure 11 indicates that this rudimentary calculation is adequate only for the limited purpose of this work. Data for $2\,h^{-1}\mathrm{Mpc} \leq s \leq 8\,h^{-1}\mathrm{Mpc}$, where s is the redshift distance (h is the Hubble parameter, $H_0 = 100\,h\,(\mathrm{km/s})/\mathrm{Mpc}$), were fitted to the power

City Group	City-City Distance from $\xi(s)$ in miles	Pixel Measure of Average City-City Distance (0.5 mi/pixel) in miles	Pixel Measure of Average City Radius in miles	$\xi(s) = \left(\frac{s}{s_0}\right)^{-\gamma}$
Omaha, Des Moines, Cedar Rapids, Davenport, Kansas City, St. Louis, Springfield	242 ± 6	228 ± 82	20 ± 5	$\left(\frac{s}{20\pm10\ \mathrm{mi}}\right)^{-0.6\pm0.2}$
Oklahoma City, Tulsa, Fayetteville, Little Rock, Dallas, Shreveport	263 ± 6	215 ± 68	22 ± 8	$\left(\frac{s}{24^{+39}_{-12}\ \mathrm{mi}}\right)^{-0.7\pm0.3}$
Nashville, Knoxville, Huntsville, Chattanooga, Birmingham, Atlanta, Montgomery	149 ± 4	155 ± 58	14 ± 8	$\left(\frac{s}{15\pm5\ \mathrm{mi}}\right)^{-0.9\pm0.2}$

TABLE I. City distances and radii as determined from the correlation function, compared to direct pixel measurements converted to miles.

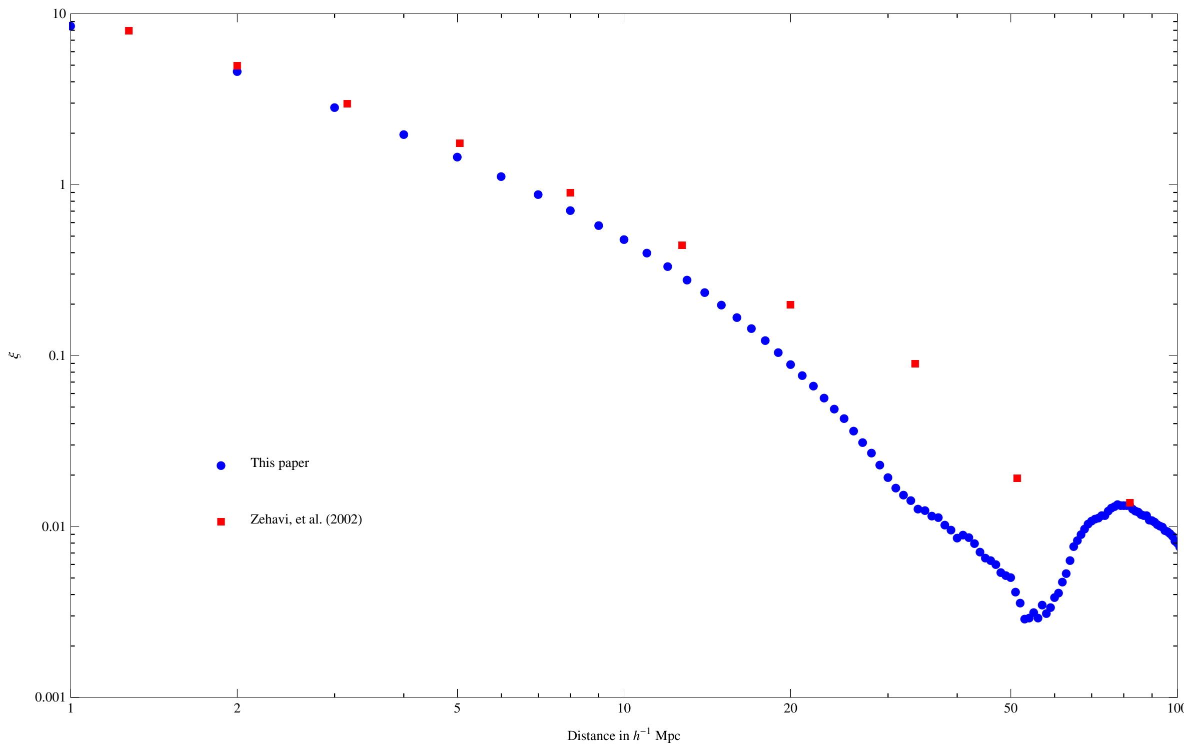


FIG. 10. Two-point correlation function for SDSS DR7 data (330,393 galaxies) compared to SDSS EDR analysis (29,300 galaxies) by Zehavi *et al.*[15]

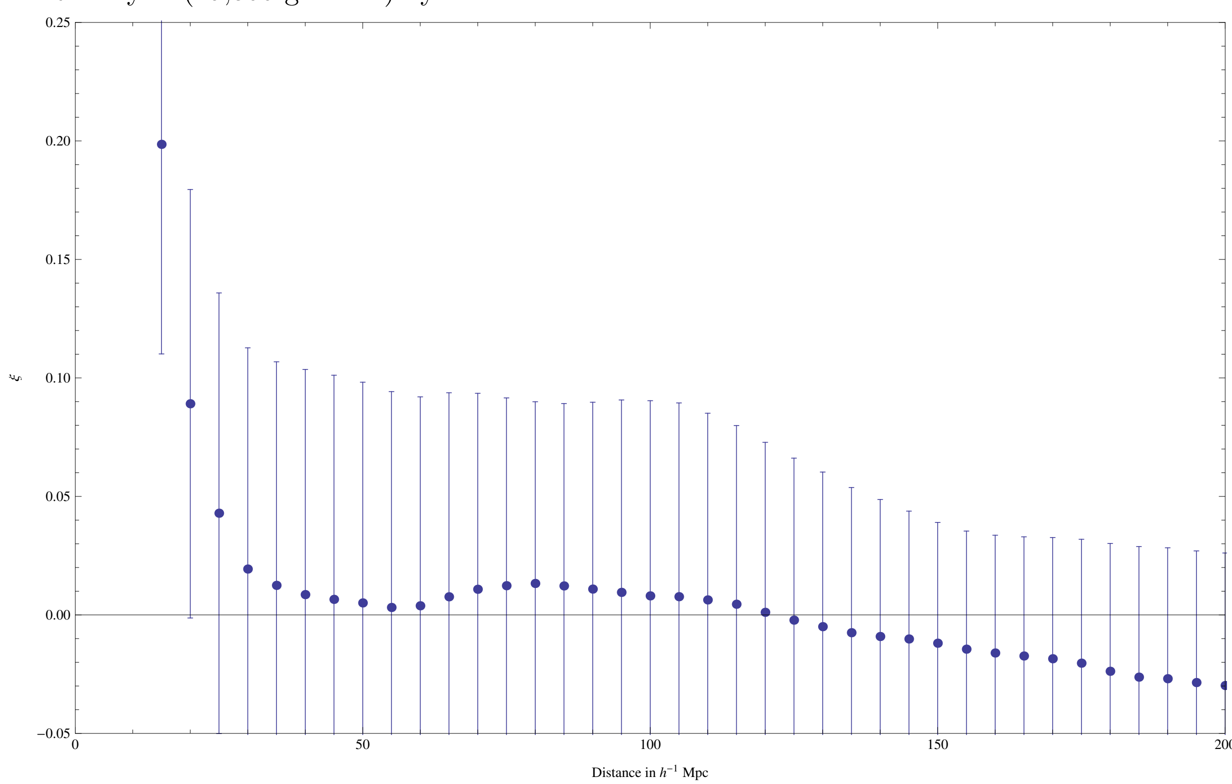


FIG. 11. 1σ errors in the galaxy-galaxy two-point correlation function.

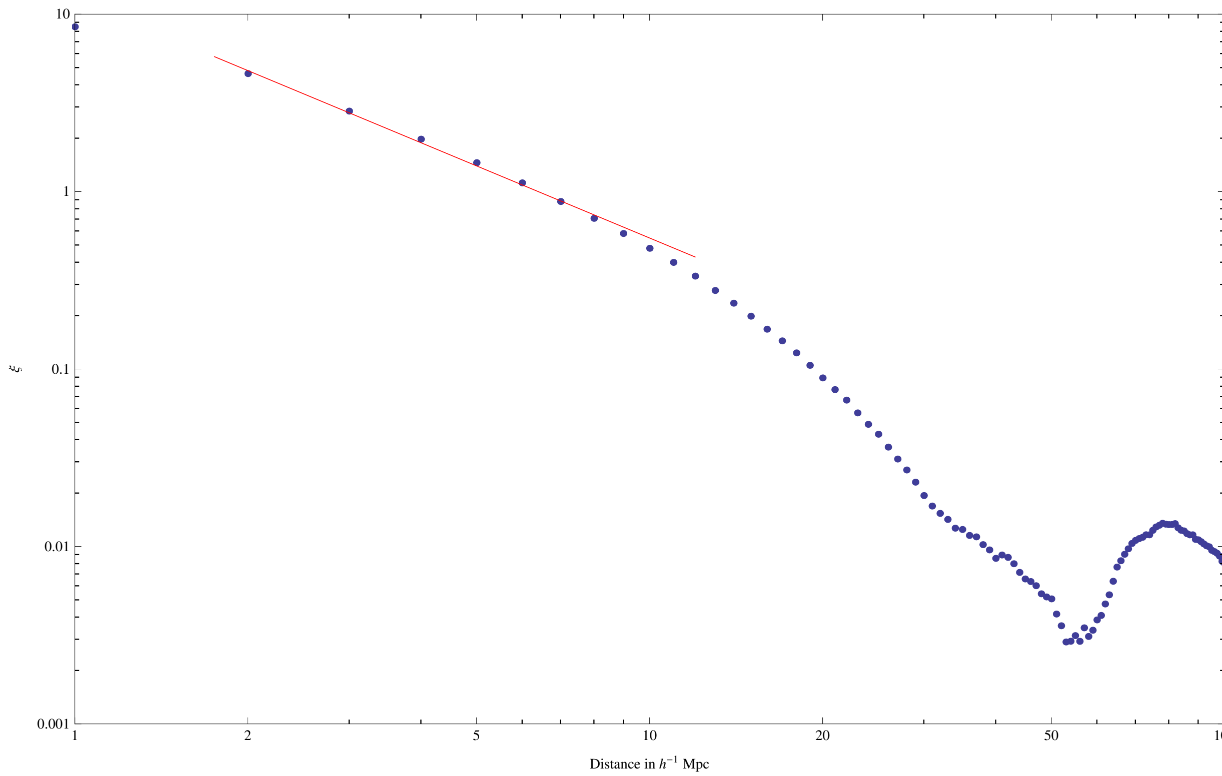


FIG. 12. Power-law fit to galaxy-galaxy correlation function: $\xi(s) = (s/6.4 \pm 0.6\, h^{-1}\text{Mpc})^{-1.4\pm0.2}$.

law, resulting in $\xi(s) = (s/6.4 \pm 0.6\, h^{-1}\text{Mpc})^{-1.4\pm0.2}$ (Figure 12), in fair agreement with $\xi(s) = (s/8.0\, h^{-1}\text{Mpc})^{-1.2}$ calculated in Ref. 15. The power law enables an interpretation based on what has been previously discovered for cities. Accordingly, the characteristic distance scale $6.4\, h^{-1}$Mpc is the radius of a galaxy cluster. For the current best value of h,[19] $h = 0.70$, this implies a galaxy cluster radius of approximately 9 Mpc. Of course this value will differ from other values derived by using different definitions of galaxy cluster.

Continuing the analogy with cities, the prominent peak in Figures 10–12 suggests a measurement of average galaxy cluster separation distance. This peak hints at the imprint of the baryons that have propagated, like a sound wave, out of dense regions of dark matter because of baryons coupling to photons—the so-called baryon acoustic peak.[20,21] Baryon-photon decoupling occurs 380,000 years after the Big Bang and results in correlations between dense dark matter regions and dense baryon regions that are manifested in galaxy distributions. (Eventually the notion of two distinct regions is erased because the dense baryon region attracts dark matter.) The naïve analysis of this work gives an average galaxy cluster separation of $s = 78 \pm 1\, h^{-1}$Mpc. But the careful analysis of Eisenstein *et al.*,[22] followed by other researchers,[23–27] actually puts the baryon acoustic peak at $s = 100\,h^{-1}$Mpc. The large discrepancy is likely due to corrections omitted, mentioned above, and because a proper analysis requires a larger volume, more easily surveyed by Luminous Red Galaxies (LRG).[22]

VI. STUDENT LABORATORY

Some aspects of this work have been adapted to a lab in introductory astronomy.[28] Students are first introduced to astronomical distances by having them consult reputable websites for typical distances to stars, and to galaxies. They are led to determine that galaxies are roughly a million times farther away than stars. Next, students are given the appropriate Structured Query Language (SQL) for downloading SDSS data ($0 < z < 0.05$, $-3° \leq dec \leq 3°$, $0 \leq ra \leq 360°$) sufficient for a wedge plot (Figure I) in a spreadsheet, about 9,600 galaxies. This enables a visual comparison between the students' galaxy clustering plot and city clustering (images provided by the instructor). Finally, students are given the data for the Midwest cities that enable them to plot the two-point correlation function so that it can be compared to the correlation function for galaxies provided by the instructor.

VII. SUMMARY

Two-point correlation functions are calculated for night satellite images of three groups of U.S. cities, and the results compared with direct pixel measurements of these cities (converted to miles) to provide an analog for the interpretation of galaxy-galaxy correlation functions. Also, the two-point correlation function is approximated for SDSS DR7 galaxies in the redshift range $0.019 \leq z \leq 0.13$. For cities and galaxies, the correlation functions can be fitted to a power law of the form $\xi(s) = (s/s_0)^{-\gamma}$. For cities, the characteristic distance, s_0, is discovered to equal the average digitally measured radius of a group of cities. Then for galaxies, s_0 is interpreted to be the radius of a galaxy cluster, approximately 9 Mpc in the ΛCDM model with $h = 0.70$.

The correlation function for each group of cities decreases from positive to negative as the distance from the city center, s, increases, then returns to a slightly positive correlation as s increases further. The value of s at the maximum of this slightly positive correlation is approximately equal to the average measured distance between all of the cities of a group.

For galaxies, there is a similar decrease in the correlation function as the redshift distance, s, increases, followed by an increase that also reaches a maximum, giving the distance between galaxy clusters. For the crudely calculated correlation function of this work, the maximum is at $s = 78\,h^{-1}$Mpc, but the thorough analysis of Eisenstein *et al.*[22] reveals that the maximum is the baryon acoustic peak with an actual value of $100\,h^{-1}$Mpc.

Three aspects of the galaxy-galaxy correlation function, therefore, have city-city analogs: the power law, the characteristic distance scale, and the city-to-city distance. This analogy should be helpful for introducing non-experts to the two-point correlation function.

ACKNOWLEDGMENTS

Support for this work has been provided by the National Science Foundation (NSF) Partnerships in Astronomy and Astrophysics Research and Education (PAARE) award AST–0750814.

---

[1] B. Ryden, *Introduction to Cosmology* (Addison Wesley, San Francisco, CA, 2003).

[2] P.J.E. Peebles, *The Large-Scale Structure of the Universe* (Princeton University Press, Princeton, NJ, 1980).

[3] J.A. Peacock *et al.*, “A Measurement of the the Cosmological Mass Density from Clustering in the 2dF Redshift Survey,” *Nature* **410**, 169–173 (2001).

[4] E. Hawkins *et al.*, “The 2dF Galaxy Redshift Survey: Correlation Functions, Peculiar Velocities and the Matter Density of the Universe,” *Monthly Notices of the Royal Astronomical Society* **346**, 78–96 (2003).

[5] D.J. Eisenstein *et al.*, “Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies,” *The Astrophysical Journal* **633**, 560–574 (2005).

[6] M. Davis and P.J.E. Peebles, “A Survey of Galaxy Redshifts. V. The Two-Point Position and Velocity Correlations,” *The Astrophysical Journal* **267**, 465–482 (1983).

[7] S.D. Landy and A.S. Szalay, “Bias and Variance of Angular Correlation Functions,” *The Astrophysical Journal* **412**, 64–71 (1993).

[8] A.J.S. Hamilton, “Toward Better Ways to Measure the Galaxy Correlation Function,” *The*

Astrophysical Journal **417**, 19–35 (1993).

[9] M. Kerscher *et al.*, "A Comparison of Estimators for the Two-Point Correlation Function," *The Astrophysical Journal* **535**, L13–L16 (2000).

[10] B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987).

[11] P. Diaconis and B. Efron, "Computer-Intensive Methods in Statistics," *Scientific American* **248**, 116–130 (1983).

[12] W.T. Sullivan III, *North America at Night* (Hansen Planetarium, Salt Lake City, UT, 1993). An updated poster is at <www.ngdc.noaa.gov/dmsp/night_light_posters.html>.

[13] *Mathematica*, Version 8.0, (Wolfram Research, Inc., Champaign, IL, 2010).

[14] IERS Technical Note No. 36, edited by G. Petit and B. Luzum (International Earth Rotation and Reference Systems Service, Frankfurt, Germany, 2010), <www.iers.org>.

[15] I. Zehavi *et al.*, "Galaxy Clustering in Early Sloan Digital Sky Survey Redshift Data," *The Astrophysical Journal* **571**, 172–190 (2002).

[16] K.N. Abazajian *et al.*, "The Seventh Data Release of the Sloan Digital Sky Survey," *The Astrophysical Journal Supplement Series* **182**, 543–558 (2009).

[17] D.W. Hogg, "Distance Measures in Cosmology," <xxx.lanl.gov/ps/astro-ph/9905116> (2000).

[18] I.V. Chilingarian, A. Melchior, and I.Y. Zolotukhin, "Analytical Approximations of K-Corrections in Optical and Near-Infrared Bands," *Monthly Notices of the Royal Astronomical Society* **405**, 1409–1420 (2010).

[19] N. Jarosik *et al.*, "Seven-Year *Wilkinson Microwave Anisotropy Probe (WMAP)* Observations: Sky Maps, Systematic Errors, and Basic Results," *The Astrophysical Journal Supplement Series* **192**, 14–28 (2011).

[20] D.J. Eisenstein and W. Hu, "Baryonic Features in the Matter Transfer Function," *The Astrophysical Journal* **496**, 605–614 (1998).

[21] S. Bashinsky and E. Bertschinger, "Position-Space Description of the Cosmic Microwave Background and Its Temperature Correlation Function," *Physical Review Letters* **87**, 1301–1304 (2001); S. Bashinsky and E. Bertschinger, "Dynamics of cosmological perturbations in position space," *Physical Review D* **65**, 3008–3026 (2002).

[22] D.J. Eisenstein *et al.*, "Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies," *The Astrophysical Journal* **633**, 560–574 (2005).

[23] E.A. Kazin *et al.*, "The Baryonic Acoustic Feature and Large-Scale Clustering in the SDSS LRG Sample," *The Astrophysical Journal* **710**, 1444–1461 (2010).

[24] V.J. Martinez *et al.*, "Reliability of the Detection of the Baryon Acoustic Peak," *The Astrophysical Journal Letters* **696**, L93–L97 (2009).

[25] C. Blake *et al.*, "The WiggleZ Dark Energy Survey: Mapping the Distance-Redshift Relation with Baryon Acoustic Oscillations," *Monthly Notices of the Royal Astronomical Society* **418**, 1707–1724 (2011).

[26] W.J. Percival *et al.*, "Baryon Acoustic Oscillations in the Sloan Digital Sky Survey Data Release 7 Galaxy Sample," *Monthly Notices of the Royal Astronomical Society* **401**, 2148–2168 (2010).

[27] A. Cabre and E. Gaztanaga, "Clustering of Luminous Red Galaxies I: Large Scale Redshift Space Distortions," *Monthly Notices of the Royal Astronomical Society* **393**, 1183–1208 (2009).

[28] <physics.scsu.edu/%7Edms/cosmology/home2.html>.