Spatial point processes in astronomy

Gutti Jogesh Babu a,*, Eric D. Feigelson b

a Department of Statistics, 319 Classroom, Blg., Pennsylvania State University, University Park, PA 16802-2111, USA
b Department of Astronomy & Astrophysics, Pennsylvania State University, University Park, PA 16802, USA

Received 17 February 1994; revised 19 September 1994

Abstract

We review several topics arising in astronomical research where the statistical analysis of spatial point processes plays a central role in scientific investigations. These include: the distribution of stars in the Milky Way; the distribution of galaxies in two and three dimensions; the location of clusters of photons from faint astronomical sources in detectors; the classification of objects in large multivariate databases; and the sky distribution of gamma-ray bursters. A variety of challenging methodological issues arise including anisotropic filamentary clustering, modeling of Poisson spatial processes, and weighted measures of clustering.

AMS Subject Classifications: Primary 85A35; Secondary 62M30, 85A05

Keywords: Anisotropy; Borel distribution; CCD detectors; Censoring; Galaxies; Gamma-ray bursters; Gravitational clustering; Infinitely divisible distributions; IRAS; Isotropic process; k-mean clustering; Malmquist bias; n-point correlation function; Point spread function; Poisson process; ROSAT; Selection bias; X-ray astronomy

1. Introduction

On a clear night, one can look upward and see the sky speckled with a complex distribution of stars and a diffuse band, the Milky Way. Galileo first turned a telescope towards the heavens in the early 17th century, and found the Milky Way is comprised of inhomogeneous distributions of countless faint stars. With larger telescopes, astronomers discovered beautiful spiral and elliptical galaxies distinct from our Milky Way galaxy, each of which contains $10^6-10^{12}$ stars like our Sun. The distribution of galaxies exhibits complex hierarchies of anisotropic clustering. The nighttime sky in visible light is thus a laboratory ready for the quantitative study of the distribution of points of light on a dark sphere or in three-dimensional space.

1 Research supported in part by NSF grant DMS-9208066 and NASA grant NAGW-2120.
* Corresponding author.
We review here five distinct areas of contemporary astronomy where the statistical analysis of spatial point processes is an integral element in the pursuit of astronomical questions. The distribution of stars within our Galaxy and the distribution of galaxies themselves are discussed first. We then turn to the distribution of photons in modern astronomical detectors, the delineation of stellar populations from multivariate data sets, and the spatial distribution of the enigmatic gamma-ray burst sources in the sky. These rather different fields, which do not at all exhaust the range of relevant problems, are chosen to illustrate the wide variety of topics where spatial point processes directly impacts astronomical understanding. We hope to leave the reader tantalized rather than satisfied, and interested in pursuing the many issues arising in astronomical spatial statistics.

2. Constellations and stellar statistics

The study of the spatial patterns of stars in the sky has been pursued, generally without mathematics, for millennia in many societies. From the perspective of modern astronomy, these efforts have had little meaning. The familiar constellations of bright stars, such as Orion, the Big Dipper or Southern Cross, in most cases are simply chance alignments of unrelated stars at different distances. Stars have an enormous range of intrinsic luminosities (from $10^{-4}$ to $10^{14}$ the luminosity of our Sun), and can have similar apparent brightnesses although they lie at very different distances from the Earth. With a few notable exceptions (such as the Pleiades or Seven Sisters, a cluster of stars formed about 50 million years ago), no significant scientific result has emerged from the study of constellation patterns.

The distribution of the more numerous fainter stars, however, was thought to be a key to the distribution of mass in the universe. For three centuries after Galileo’s discovery that the Milky Way is comprised of myriad faint stars, it was hoped that mapping the surface density of stars in different directions would lead to a reliable determination of the structure of the galaxy (Herschel 1785; or see the reprint in Hoskin, 1964, pp. 82–106; Kapteyn, 1922). However, by the 1920–1930s, it was recognized that the space between the stars in the Milky Way is not empty, but rather it contains a very inhomogeneous distribution of clouds of dust and gas. These dark interstellar clouds obscure light from stars behind them, so stellar surface densities do not directly measure distances. Due to this interstellar obscuration, it has proved difficult or impossible to invert the ‘fundamental equation of stellar statistics’ (Mihalas and Binney, 1981) and derive the shape of the galaxy. This long effort to derive global structure of the galaxy from stellar statistics is documented by Trumpler and Weaver (1953, §5).

3. Spatial distribution of galaxies

Edwin Hubble was the founder of extragalactic astronomy, where the objects of study are distinct galaxies rather than individual stars within our own Galaxy.
his many achievements, Hubble (1934) made the first quantitative study of the spatial frequency distribution of galaxies in different directions. He reported that the count $N$ of galaxies in a telescope field is skewed from the normal distribution, resembling rather a lognormal. When more thorough surveys of galaxy counts became available, it became clear that galaxies were clustered. Berkeley statisticians Neyman and Scott (1952, see review in Neyman, 1962) adopted a double-Poisson clustering model where the cluster centers are randomly distributed.

### 3.1. $n$-point correlation functions

This statistical approach to galaxy clustering (in contrast to attempts to locate individual groups or clusters directly on the sky; e.g. Abell, 1958; Turner and Gott, 1976) developed considerably during the 1970s. Peebles (1980, and references therein) and others calculated the $n$-point correlation functions to describe the distribution of galaxies in the universe. The second moment of the number of objects in a region depends only on the two-point correlation, the third moment depends on the three-point correlation, and so on. Astronomers attempt to map the distribution of matter in the universe from the distribution of luminous point-like galaxies or clusters of galaxies. Large catalogs giving galaxy positions and redshifts (which is linearly proportional to distance, with some scatter) are available.

If differences among the objects are ignored, the distribution can be described entirely in terms of positions of these objects, which can be described in terms of $n$-point correlation functions. This function is defined as follows. If the galaxies are distributed uniformly in a region of the sky, then the probability density of the distribution of the matter in the region is a constant, $m$. Let $f$ denote the probability density of the distribution of galaxies within a circular region $V$ around a given galaxy. If the universe is assumed to be both homogeneous and isotropic, on a large scale, and the distribution of objects close by are correlated, then $f$ may assume the form

$$f(V) = m(1 + \xi(V)),$$

where the so-called correlation function $\xi$ depends only on the distance $r$ from the reference galaxies. If the integral

$$m \int \xi \, dV = n_c - 1$$

converges, then $n_c$ can be interpreted as a measure of the mean number of objects per cluster. The integral gives the total number of objects in excess of those predicted using uniform distribution.

Consider, for example, a simplified universe where all galaxies reside in clusters each with diameter $D$ and each containing $n_c$ members, and that the cluster centers are uniformly distributed. A galaxy chosen at random also identifies the cluster to which it belongs. As the clusters are randomly placed, they contribute $mV$ to the average, same
as for randomly placed volume. In addition, there are \( n_c - 1 \) neighbors from the chosen cluster, so the integral of \( m\xi \) above is \( n_c - 1 \). If the number of objects per cluster is instead a random variable, the probability of choosing an object from a cluster with \( n_a \) members is proportional to \( n_a \), leading to the equation

\[
m \int \xi \, dV = E(n_a(n_a - 1))/E(n_a),
\]

where the expectation is taken over the distribution of clusters in space. The two-point correlation \( \xi \) is defined through the probability density

\[
f(V_1, V_2) = m^2(1 + \xi(V_1, V_2))
\]

governing the existence of galaxy pairs in regions \((V_1, V_2)\). In view of the assumed isotropic and homogeneous cosmological model, \( \xi(V_1, V_2) = w(\theta) \) depends only on the angular distance \( \theta \) between the regions \( V_1 \) and \( V_2 \).

Suppose the joint density describing the probability of finding objects in each of the three regions \( V_1, V_2, V_3 \) is given by

\[
f(V_1, V_2, V_3) = m^3[1 + w(r_a) + w(r_b) + w(r_c) + \xi(r_a, r_b, r_c)], \tag{1}
\]

where \( r_a, r_b, r_c \) are the sides of the triangle defined by the three points. Then \( f(V_1, V_2, V_3)/m^3 \) is called the full three-point correlation function, and \( \xi \) is called the reduced three-point correlation function. If \( V_1 \) and \( V_2 \) are close by and \( V_3 \) is far away, then the chance of finding an object in it is unaffected by what happens in the first two. Consequently, \( \xi \) vanishes at large \( r \), and one can conveniently treat the reduced function \( \xi \) as a perturbation. In Eq. (1), the dominant term in the square brackets may be the first, which represents triplets of galaxies at three very different distances accidentally seen close together in the sky. The next largest terms are the three two-point functions that represent a pair of galaxies close together in space with the triplet accidentally completed by a third galaxy at a very different distance. The form of the equation conveniently separates accidental and physical triplets in the data.

The moment restrictions impose certain restrictions on \( w \) and \( \xi \). The observed distribution of large galaxy samples has been extensively measured and modeled using \( n \)-point correlation functions (Peebles, 1980). The resulting good approximations for two- and three-point correlation functions are

\[
w(\theta) = B\theta^{-\gamma}, \quad \gamma \approx 1.77,
\]

\[
\xi(r_a, r_b, r_c) = Q[w(r_a)w(r_b) + w(r_b)w(r_c) + w(r_c)w(r_a)],
\]

where \( Q \) is a constant. Other models suggested include the Gaussian model, where \( w(\theta) = A \exp(-[(\theta/\theta_0)^2]) \).

Although \( w \) is not a correlation function in the usual statistical sense, in some instances it is estimated as an auto-correlation based on replicated observations or superimposed data from similar regions (see Peebles 1980, §33). The auto-correlation
function \( w \) is certainly not the only statistic relevant to galaxy clustering. It does not accurately characterize the abundance of rare extreme fluctuations like the Abell clusters because it is not very sensitive to them. The model becomes complicated, once the effect of redshift is also included in the model, in addition to the galaxy positions in the celestial sphere (Peebles, 1980, Ch. 4).

All of these statistical descriptions of the galaxy distribution assume that the brighter galaxies near our Milky Way constitute a fair sample of the Universe. However, the assumption of unbiased sampling in galaxy surveys can not be adopted without considerable thought. Selection biases in the data can include: obscuration by dark clouds in our own Galaxy; obscuration by dark clouds in the other galaxies; the 'Malmquist' or truncation bias due to the magnitude limit of galaxy surveys; morphological segregation of galaxies in clustered environments; cosmic evolution of galaxy luminosities and colors; and more. Statisticians interested in galaxy clustering are encouraged to work with extragalactic astronomers who are familiar with these limitations in the data sets.

### 3.2. Saslaw's galaxy distribution function

One astrophysical model of galaxy clustering that has emerged with a clear prediction for the spatial distribution of galaxies today has been developed by Saslaw (1985). Based on a thermodynamic model of gravitational clustering, it predicts that the probability \( f(N) \) for finding \( N \) galaxies in a three-dimensional space \( V \) with volume \( v \) is

\[
f(N) = \frac{\bar{N}(1 - b)}{N!} [\bar{N}(1 - b) + Nb]^{N-1} \exp\{-\bar{N}(1 - b) - Nb\},
\]

where \( \bar{N} = \bar{n}v \), \( \bar{n} \) is the mean number density of galaxies, and \( b = -W/2K \) is the only free parameter of the model. Here, \( W \) is the gravitational potential energy and \( K \) is the kinetic energy of peculiar motions emerging from the Big Bang.

This distribution function appears to be relatively unstudied. It is infinitely divisible and has the generating function (Saslaw, 1989)

\[
g(s) = \sum_{N} f(N) s^N = \exp\{[-\bar{N} + \bar{N}h(s)]\},
\]

\[
h(s) = b + \frac{(1 - b)}{b} \sum_{N=1}^{\infty} \frac{N^{N-1}}{N!} b^N \exp\{-Nb\} s^N.
\]

Saslaw's function (2) reduces to the Poisson distribution for \( b = 0 \), i.e. no gravitational interaction. For \( b > 0 \), it is a compound Poisson distribution. The form of \( g \) in Eq. (3) indicates that

\[
N = \begin{cases} 
0 & \text{if } M = 0, \\
X_1 + \cdots + X_M & \text{if } M > 0,
\end{cases}
\]
where \( N, X_1, X_2, \ldots \) are independent, the distribution of \( M \) is Poisson, and the probability distribution \( t \) of \( X_1 \) is given by

\[
t(n) = \begin{cases} 
  b & \text{if } n = 0, \\
  (1 - b)^{n-1} b^{n-1} \frac{n!}{n!} \exp\{-nb\} & \text{if } n > 0.
\end{cases}
\]

(4)

In this sense, the distribution of \( N \) can be interpreted as a distribution of intermingled clusters whose centers have a Poisson spatial distribution and whose probability \( t(n) \) for containing \( n \) galaxies is given by the generating function \( h(s) \). Thus we have a decomposition of \( N \). This process of model building is in reverse to what is common practice in galaxy clustering. Models for galaxy clustering are, generally, made by assuming subcluster centers to be randomly distributed and by making assumptions on the probability distribution of the number of galaxies in a subcluster. The Borel distribution (4) has been studied in detail in statistical literature in connection with queues.

From Eq. (3), the number \( N_V \) of galaxies in a region \( V \) can be treated as an integer valued stochastic process indexed by subsets of \( \mathcal{P}^3 \) with the property that \( N_V \) and \( N_U \) are independent, whenever \( U \) and \( V \) are disjoint regions. Eq. (2) has been extensively compared with observed galaxy clustering and \( N \)-body simulations of gravitational clustering in an expanding universe (e.g. Itoh et al., 1993). Good fits are obtained with \( b = 0.75 \pm 0.05 \).

3.3. Three-dimensional galaxy distributions

The statistical descriptions of the distribution of galaxies given above assume that the brighter galaxies near our Milky Way constitute a fair sample of the entire Universe, and that any reasonable astrophysical cosmological model giving rise to clustering should be a homogeneous and isotropic process. Furthermore, the samples from regions separated far from each other should be uncorrelated. These assumptions seem quite reasonable, since clustering is an expected consequence of the mutual gravitational attraction of galaxies, which obeys an isotropic inverse-square law. The spatial distribution of nearby galaxies should thus be the result of a random process and, if the galaxies are selected without bias, they should constitute an independent and identically distributed sample.

However, extensive new data have emerged over the past decade suggesting that the three-dimensional galaxy distribution has considerable anisotropy. In addition to measuring the two angular coordinates of galaxies on the celestial sphere, one can measure their velocity of motion away from us. Measuring redward Doppler shifts in galaxy spectra, Hubble discovered in the 1920s that galaxies in all directions are receding from us with a recessional velocity proportional to their distance. Known as Hubble’s Law, the only reasonable explanation is that the Universe is expanding from a compressed state some 10–20 billion years ago. Aside for the implications for cosmology, measuring galaxy redshifts gives a third dimension (two angular coordinates...
plus a velocity coordinate) to galaxy locations. Tens of thousands of nearby galaxy redshifts are now known (Giovanelli and Haynes, 1991), and measurements of $10^5$–$10^6$ galaxy redshifts are underway.

The surprise is that the clustering of galaxies in three dimensions is highly anisotropic. Rather than appearing as spherical concentrations in a smooth background, galaxies collect along the edges of giant empty spheres (de Lapparent et al., 1986). The structures, which can be considerably larger than the traditional galaxy clusters (Abell, 1958), are variously called superclusters, sheets or filaments, voids, Great Walls or Great Attractors. This large-scale filamentary superclustering was not expected, and it is proving difficult to explain it within the context of contemporary cosmological theory. Fig. 1 shows the radial distribution of galaxies in the region of the main ridge of the Pisces–Perseus supercluster.

Astronomers have begun to characterize quantitatively this new and unusual three-dimensional galaxy clustering. Recent approaches include three-dimensional versions of the n-point correlation function, power spectrum analysis, topological genus, void probability functions, Voronoi or Dirichlet tessellation, various filament-finding algorithms,
Lee statistics, percolation, minimal spanning trees, and modeling with multifractals. This exciting and extensive literature has recently been reviewed by Haynes (1992), Barrow (1992), Coles (1992) and Beers (1992). The reader is referred to these more extensive discussions.

Since the work of Neyman and Scott, few statisticians have addressed the important and complex statistical problems encountered in galaxy clustering. The problem is difficult because few existing methods are adapted to anisotropic or filamentary clustering. Perhaps methods from stochastic geometry (Stoyan et al., 1987) can be applied to galaxy clustering.

4. Sources in photon-counting detectors

An astronomical instrument typically has two elements: a telescope which focuses light onto a surface using curved lenses or (more commonly) mirrors; and a detector which records the light for scientific analysis. The oldest such instrument is the human eye, where the retina is a rather sophisticated detector that sends image information to the brain 10–20 times a second. Photographic emulsions developed in the 19th century permit exposures up to $10^3$ s, allowing the discovery and study of millions of stars and thousands of galaxies. Photographic plates proved to be rather inefficient detectors with fewer than 2% of incident light photons producing a chemical darkening of the image. In the 1970s, a series of new and remarkable detector technologies emerged from the solid state physics and micro-electronic engineering. The best of these solid-state detectors is the charged-coupled device or CCD, which can detect up to 90% of the incident light. Seen face-on, a CCD resembles a synthetic retina, with $10^5$–$10^6$ individual detecting elements or pixels. A photon hitting the surface produces a digital electrical charge at the back of the pixel. After the exposure at the telescope, the image is read into a computer. CCDs are now the preferred detectors for many applications in optical, ultraviolet and X-ray astronomy. Other detectors which detect individual energetic photons or elementary particles from stars and galaxies include imaging proportional counters, Cerenkov detectors, scintillation counters, and large tubs of water or ice. Collectively, photon-counting detectors are increasingly important in visible-light astronomy and dominate the rapidly growing fields of X-ray, gamma-ray, neutrino and cosmic ray astronomy.

The advent of photon-counting instrumentation has led to a variety of statistical problems, many of which fall under the rubric of spatial point processes. One common problem is the detection of a statistically significant source in a photon-counting device that has background noise. If the background photons $B$ are due to electronic noise, for example, they will be a random Poisson process in the image. Typically, one seeks the number of source counts $S$ from a small region within which one finds $S + B$ photons, and the expectation $\bar{B}$ is measured away from the source. If $S + B$ has $\geq 10 - 20$ photons, then Gaussian statistics can be assumed. The significance of the source, or its signal-to-noise ratio $S/N$, can then be found from tables for the Gaussian distribution,
where the estimated standard deviation is given by
\[ \sigma = S/N = \frac{S + B - \bar{B}}{\sqrt{S + 2B}}. \]

If the background is uniform, it can be measured with greater precision over a large source-free area of the detector. Tables for low count rates, where the Poisson rather than Gaussian distribution applies, are presented by Gehrels (1986).

Realistic problems, however, are often more complex. The background rate is often temporally and/or spatially variable. Temporal variations can occur because the X-ray or gamma-ray telescope resides on a satellite orbiting the Earth every \( \approx 90 \) min, traveling through different portions of the Earth's magnetosphere with variable populations of energetic particles. Spatial variations can occur because the detector construction obscured portions of the image. Fig. 2 shows an example of an image from Positional Sensitive Proportional Counter on board the German/US/UK ROSAT X-ray astronomical satellite. The expected spatial distribution of background photons is expected to be higher in the central region than near the edges, due to telescope vignetting. Elsewhere in the image, both background and source photons are obscured by aluminum rods supporting the plastic window of the detector.

![Fig. 2. A portion of a ROSAT satellite X-ray image, pointed at a nearby interstellar cloud with dozens of recently formed stars. The region shown is about the size of the full Moon. The contours represent 1, 3, 9, and 27 photons/20\( \times \) 20\( \times \) pixel. The small dots represent uninteresting background events, and the number sources represent sources with \( S/N > 3.5 \) (Feigelson et al., 1993).](image)
The ROSAT source detection problem is further complicated by variations in the telescope focusing capability across the detector. Point sources far from the detector center produce a large blur of X-ray photons, while sources close to the center produce a tightly concentrated distribution of photons. This response of a telescope–detector combination to a point source is called the point spread function. The background noise \( B \) may be viewed as a temporally and spatially variable function mixture of two components, say \( B(t,r) = C(t) + D(r) \). Here \( C(t) \) is the flux of uninteresting particles from Earth’s magnetosphere, which varies with the satellite orbit, and \( D(r) \) is the flux of real X-ray photons from many extremely faint distant X-ray sources which is attenuated off-axis by limitations in the telescope. Both components should be Poisson processes, though with mean values to be determined from source-free regions of the image. The source detection algorithms should therefore depend on location in the detector.

The statistical problem is thus to optimally locate statistically significant X-ray sources, which are Poisson clusters of photons with a known (but spatially variable) point spread function, superposed on an unclustered but variable and imprecisely quantified Poisson distribution of background photons. This problem is surprisingly important. If the statistical algorithm is too inefficient or conservative, the source flux limit is too high and many (perhaps most) of the sources in the field are missed. A later satellite costing hundreds of millions of dollars might be built to obtain greater sensitivity. If the statistical algorithm inaccurately evaluates the statistical significance of faint sources, the resulting source catalog may contain many false entries. Decades of follow-up research may be wasted tracking down these errors.

The ROSAT pipeline software implements several versions of a Poisson-based maximum-likelihood criterion proposed by Cruddace et al. (1988). Two models of the background are adopted, one based on a prediction of the background based on knowledge of the satellite’s orbit, and another on actual local background counts found around each potential source. The source list obtained by these methods is reasonably successful, but often either misses some weak sources or includes some statistically insignificant sources.

The interested reader is encouraged to read recent discussions of the statistical problem encountered in photon-counting detectors by Nousek (1992), Marshall (1992) and Bickel (1992). Nousek discusses the proposal by Kraft et al. (1991) to use a Bayesian approach to source detection confidence intervals when very few source and background counts are present. Bickel, however, argues that the assumption of a uniform Bayesian prior is unwarranted. A less controversial procedure is the maximum-likelihood \( C \) statistic based on the Poisson model recommended by Cash (1979). It is a flexible method that permits simultaneous estimates of \( S \) and \( B \), but probabilities associated with its confidence levels are not readily determined. Bootstrap resampling of photons in the image might give model-free estimates of source existence probabilities. Marshall describes the realistic problems encountered in evaluating source existence and significance for the Extreme Ultraviolet Explorer satellite. Here again, a combination of maximum-likelihood methods combined with prior knowledge of the point spread function and background level is used.
Finally, we note that the problem of identifying significant structures in low-count rate data must sometimes be extended beyond the two physical dimensions of a detector. X-ray and gamma-ray astronomical sources are frequently variable in intensity, and may thus appear statistically significant in one observation but not in another. Also, with the advent of X-ray detecting CCDs with good spectral, as well as spatial, resolution, sources can be sought in specific spectral bands or emission lines. Thus, the most general problem encountered in astronomy is source detection and characterization in four dimensions: the two dimensions of the detector (or equivalently, location on the celestial sphere), time and photon energy.

5. Stellar populations

For over a century, astronomers have devoted considerable energies dividing stars into various classification schemes to understand their physical properties and origins. The most successful system is the Harvard Spectral Classification, developed around the turn of the century by Annie Jump Cannon and her colleagues. The visible band colors and spectral lines of stars are placed in a one-dimensional sequence, OBAFGKM, where O stars are the hottest and bluest while M stars are the coolest and reddest. Later, Morgan and Keenan added a second dimension called luminosity class, indicating whether the star is a giant, supergiant, white dwarf or ordinary main sequence star (like the Sun). These classifications, often represented on the two-dimensional Hertzsprung–Russell diagram, led to the great insights of stellar structure and evolution. Stellar spectral classification is traditionally performed by the astronomer’s unaided eye–brain recognition of spectral line patterns. This method still continues today, though some researchers are trying sophisticated statistical classification procedures.

Other types of stellar classifications, however, have been more problematic than visible band spectroscopic classification. The Infrared Astronomical Satellite (IRAS), developed by a US/UK/Netherlands consortium, produced during the 1980s an extraordinary view of the sky at far-infrared wavelengths. Hundreds of thousands of infrared sources were found, almost all previously unknown. The first challenge is to separate sources associated with stars in our Galaxy from other distant galaxies. In most cases, this can be done simply by examining visible band photographs and seeing what lies in small region where the infrared source lies. But the category of stellar infrared sources proved to be quite heterogeneous. Some are young stars recently formed, or even protostars now condensing from the interstellar medium. Others are old red giants near the end of their stellar lives, ejecting dense winds of dusty gas which emit copiously in the infrared. Others, surprisingly, were found to be ordinary main sequence A stars which are infrared emitting dusty disks. In some cases, the infrared sources proved not to be stars at all, but rather clumps of cold dusty interstellar gas.

Each of these categories of IRAS sources has characteristic infrared spectral signatures and spatial distributions in the sky. The satellite had detectors in four bands centered on wavelengths of 12, 25, 60 and 100 μm. Thus, the IRAS source databases,
the Point Source Catalog and Faint Source Catalog (Moshir et al., 1992), consist of $10^5$ objects with brightness measurements at four wavebands and sky location measurements in two-dimensions. Fig. 3 shows an example of the complexity of the databases, in a plot of the ratios of the 25-μm/12-μm and 60-μm/100-μm. In this database, one needs to consider spatial statistical methods in six dimensions: two locations on the celestial sphere and four spectral bands.

This complex but highly organized distribution in six-dimensional space is amenable to modern statistical analysis and classification based on $k$-mean clustering. The $k$-mean $(\mu_1, \ldots, \mu_k)$ based on the sample $X_1, \ldots, X_n$, is the vector which minimizes

$$W_n(a_1, \ldots, a_k) = \sum_{i=1}^{n} \min_{1 \leq j \leq k} (X_i - a_j)^2.$$  

This gives a classification based on those $X_i$'s that are closest to each of the $\mu_j$. Its asymptotic theory is presented by Hartigan (1978), Pollard (1981, 1982) and Serinko and Babu (1992, 1994). Interestingly, in many situations it turns out that the behavior at large-$n$ is non-Gaussian. In the double exponential case, the limiting distribution of 2-means is concentrated on the line $x = y$, and that the marginal distribution is non-normal. In fact, Serinko and Babu (1992) have shown that the limit of the marginal distribution is given by $a \text{sign}(W) \sqrt{|W|}$, and $W \sim N(0, 1)$, and $a$ is a constant. The rate of convergence is $n^{1/4}$. The limiting structure seems to depend strongly on smoothness of the density at $(\mu_1, \ldots, \mu_k)$. Murtagh (1992) presents results from a $k$-means
partitioning of three color–color ratios from the IRAS Point Source Catalog, projected onto the two spatial dimensions. The results show that the structures seen in the color–color plots like Fig. 3 isolate distinct populations in the galaxy.

So far only a handful of studies have been made of the IRAS and other large-scale astronomical databases using multivariate classification methods. Such problems do not only arise in surveys from satellite-borne observatories. The problem of discriminating clusters and patterns of points in multidimensional spaces occurs frequently in the field of galactic astronomy. Perhaps the most famous example was the discovery by J. Oort in the 1920s of a halo population of stars in the solar neighborhood, in addition to the dominant galactic disk stellar population. The two populations form different asymmetrical distributions in three-dimensional plots of stellar velocity vectors (velocity ellipsoid diagrams, see Mihalas and Binney 1981, Ch. 7). Today, while the existence of galactic halo stars is established, controversy continues over whether a third stellar component, the ‘thick disk’, is present in the galaxy (Gilmore et al., 1989). The statistical problem is to determine whether distinct clusters are present in asymmetrical distributions of points in multidimensional spaces that include stellar positional, kinematic, metallicity and spectral variables. Enormous astronomical surveys with $10^6$–$10^8$ objects are emerging during the 1990s, which will generate many problems in the analysis of multivariate spatial databases.

The principal difficulties posed by astronomical databases of this type are due to measurement errors and censoring. As described in Section 4, sources are identified by the presence of a significantly high $S/N$ ratio at a given sky location. In the IRAS all-sky survey, it is quite common that a star is detected in one or two infrared bands, but not all four bands. Only an upper limit, or censored value, is available at the other bands and the color ratios are also censored consequently (Feigelson, 1992). That is, the observed vectors $Y = (y_{i1}, \ldots, y_{ik})$ and the actual characteristics of a star $X = (x_{i1}, \ldots, x_{ik})$ are related by $y_{ij} = \max(x_{ij}, c_{ij})$ in the censored case, where $c_{ij}$ are censoring variables. In the case of truncation due to limited sensitivity of the instruments, the observations may be recorded only when $x_{ij}$ is less than some known $t_{ij}$. Even sources detected at all bands are subject to measurement errors, $Y = X + \eta$, where the variance of the error $\eta$ is known from the original measurements of $S$, $B$ and $S/N$. Thus, each point in Figure 3 has an associated known standard deviation along each ratio axis, and these errors differ from point to point. Furthermore, points censored in one or both axes exist and should be added to the analysis.

The statistical challenge is to generalize existing spatial point process clustering and multivariate classification models, such as discriminant analysis and $k$-means partitioning, to include effects of known heteroscedastic measurement errors, selection biases such as truncation, and possible censoring in each variable. The situation is helped by the astronomers’ prior knowledge. Prototypes of each prospective class are already known and well-studied, so classification can be ‘supervised’. Perhaps neural network or Bayesian classifier methods can supplement other multivariate statistical procedures. The interested reader is encouraged to examine the discussion by Murtagh (1992).
6. Gamma-ray bursters

One of the most exotic and mysterious phenomena in modern astronomy are the gamma-ray bursters. They were first discovered in the 1960s by a group of US satellites designed to monitor nuclear explosions in outer space prohibited by the US–Soviet Test Ban Treaty. No nuclear explosions were found, but instead occasional powerful bursts of gamma-ray emission from outside the solar system were discovered. Except for three 'soft gamma-ray repeaters' which appear to constitute a distinct class, it is not at all evident what type of astronomical object produces these energetic explosions (normal stars, collapsed stellar remnants like neutron stars or black holes) or where these objects lie (close to the Sun, in the galactic disk or halo, in distant galaxies, or at cosmological distances).

The favored model of the 1980s was some type of electromagnetic instability above the surfaces of neutron stars in the galactic disk. But the Burst And Transient Source Experiment (BATSE) on the recently-launched US Compton Gamma Ray Observatory has contradicted this model. With increased sensitivity and resolution, several hundred bursters have been located (Fig. 4). They do not cluster along the galactic plane (galactic latitude 0°), as predicted by the disk model. The preferred models now involve exotic explosions in distant galaxies, such as the collision of two neutron stars.

With a lack of astrophysical insight, much of the research into these objects has been statistical in nature. This includes comparing the numbers seen at each brightness level with model predictions, classifying the diverse temporal and spectral signatures of the bursts, and studying their spatial distribution for anisotropies and/or repeated

![Spatial distribution of 261 gamma-ray bursts observed by BATSE on the Compton Gamma-Ray Observatory. The diagram shows the entire celestial sphere, with the plane of the Milky Way oriented horizontally across the middle. Differing error boxes for each burst are not shown (Fishman et al., 1993).](image)
events. The spatial analysis is often focussed on whether the sources exhibit a dipole or quadrapole anisotropy indicating a concentration along the Milky Way (Paczyński, 1990). With new BATSE data, controversies are flourishing, for instance, whether the nearest neighbor distribution of burst locations implies some repetition (e.g. Quashnock and Lamp, 1993; Narayan and Piran, 1993). Readers interested in gamma-ray burst statistics are referred to Ho et al. (1992) and Shrader et al. (1992), and frequent recent articles in the *Astrophysical Journal* and *Monthly Notices of the Royal Statistics Society*.

References


