Research in
Astronomy and
Astrophysics

An algorithm for preferential selection of spectroscopic targets in LEGUE [*]

Jeffrey L. Carlin[1], Sébastien Lépine[2,3], Heidi Jo Newberg[1], Li-Cai Deng[4],
Timothy C. Beers[5], Yu-Qin Chen[4], Norbert Christlieb[6], Xiao-Ting Fu[4], Shuang Gao[4],
Carl J. Grillmair[7], Puragra Guhathakurta[8], Zhan-Wen Han[9], Jin-Liang Hou[10],
Hsu-Tai Lee[11], Jing Li[4], Chao Liu[4], Xiao-Wei Liu[12], Kai-Ke Pan[13], J. A. Sellwood[14],
Hong-Chi Wang[15], Fan Yang[4], Brian Yanny[16], Yue-Yang Zhang[4], Zheng Zheng[17] and
Zi Zhu[18]

[1] Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, 110
    8th Street, Troy, NY 12180, USA; *carlij@rpi.edu*
[2] Department of Astrophysics, American Museum of Natural History, Central Park West at 79th
    Street, New York, NY 10024, USA
[3] City University of New York, 365 Fifth Avenue, New York, NY 10016, USA
[4] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China
[5] National Optical Astronomy Observatory, Tucson, AZ 85719, USA
[6] Center for Astronomy, University of Heidelberg, Landessternwarte, Königstuhl 12, D-69117
    Heidelberg, Germany
[7] Spitzer Science Center, 1200 E. California Blvd., Pasadena, CA 91125, USA
[8] UCO/Lick Observatory, Department of Astronomy and Astrophysics, University of California,
    Santa Cruz, CA 95064, USA
[9] Yunnan Astronomical Observatory, Chinese Academy of Sciences, Kunming 650011, China
[10] Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China
[11] Institute of Astronomy and Astrophysics, Academia Sinica, P.O. Box 23-141, Taipei 106, China
[12] Department of Astronomy & Kavli Institute of Astronomy and Astrophysics, Peking University,
    Beijing 100875, China
[13] Apache Point Observatory, PO Box 59, Sunspot, NM 88349, USA
[14] Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road,
    Piscataway, NJ 08854-8019, USA
[15] Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210008, China
[16] Fermi National Accelerator Laboratory, P.O. Box 500, Batavia, IL 60510, USA
[17] Department of Physics and Astronomy, University of Utah, Salt Lake City, UT 84112, USA
[18] Department of Astronomy, Nanjing University, Nanjing 210011, China

Abstract We describe a general target selection algorithm that is applicable to any
survey in which the number of available candidates is much larger than the number of
objects to be observed. This routine aims to achieve a balance between a smoothly-
varying, well-understood selection function and the desire to preferentially select cer-
tain types of targets. Some target-selection examples are shown that illustrate different

possibilities of emphasis functions. Although it is generally applicable, the algorithm was developed specifically for the LAMOST Experiment for Galactic Understanding and Exploration (LEGUE) survey that will be carried out using the Chinese Guo Shou Jing Telescope. In particular, this algorithm was designed for the portion of LEGUE targeting the Galactic halo, in which we attempt to balance a variety of science goals that require stars at fainter magnitudes than can be completely sampled by LAMOST. This algorithm has been implemented for the halo portion of the LAMOST pilot survey, which began in October 2011.

Key words: surveys: LAMOST — Galaxy: halo — techniques: spectroscopic

1 INTRODUCTION

This document describes the algorithms used to select stellar targets for the Milky Way structure survey known as LEGUE (LAMOST Experiment for Galactic Understanding and Exploration). LEGUE is one component of the LAMOST Spectrocopic Survey (see Zhao et al. 2012 for an overview) that will be carried out on the Chinese Guo Shou Jing Telescope (GSJT). The GSJT has a large (3.6–4.9-meter, depending on the direction of pointing) aperture and a focal plane populated with 4000 robotically-positioned fibers that feed 16 separate spectrographs, providing the opportunity to efficiently survey large sky areas to relatively faint magnitudes.

The motivation for this algorithm was a desire for a well-understood and reproducible selection function that will enable statistical studies of Galactic structure. A continuous selection function is desirable, rather than assigning targets by, for example, ranges in photometric color, and excluding targets outside the color-selection ranges. Another motivation for this scheme was the opportunity provided by the sheer scale of the planned LAMOST survey; the possibility of observing a large fraction of the available Galactic stars (at high latitudes, at least) along any given line of sight allows for less stringently-defined target categories, since a more general selection scheme can gather (nearly) all of the stars in particular target categories, while simultaneously sampling all other regions of parameter space. This opens up a large serendipitous discovery space while also enabling studies of all components of the Milky Way.

The LEGUE survey will obtain an unprecedented catalog of millions of stellar spectra to relatively faint magnitudes (to at least 19th magnitude in the SDSS r-band) covering a large contiguous area of sky. The only large-scale spectroscopic survey of comparable depth is the Sloan Extension for Galactic Understanding and Exploration (SEGUE; Yanny et al. 2009a), which has been an enormously valuable resource for studies of Milky Way structure (e.g., Allende Prieto et al. 2006; Carollo et al. 2007; Xue et al. 2008; Dierickx et al. 2010; Chen et al. 2011; Lee et al. 2011; Cheng et al. 2012; Smith et al. 2012) and substructure (see Newberg et al. 2002; Yanny et al. 2003; Belokurov et al. 2006; Grillmair & Dionatos 2006; Newberg et al. 2007; Klement et al. 2009; Schlaufman et al. 2009; Smith et al. 2009; Yanny et al. 2009b; Xue et al. 2011)[1]. However, the SEGUE survey was limited to $\sim 300\,000$ stellar spectra in ~ 600 separate 7 square degree plates spread over the Sloan Digital Sky Survey Data Release 8 (DR8; Aihara et al. 2011) footprint. The separation of SEGUE into discrete "plates," while providing sparse coverage of all of the Galactic components (as well as sampling a number of known substructures), creates some difficulty in interpreting results from SEGUE. In addition, the limited number of targets observed by SEGUE necessitated selecting small numbers of stars from carefully defined target categories, most of which were delineated by selections in photometric color (Yanny et al. 2009a). This "patchy," non-uniform selection function makes statistical studies of Galactic structures difficult. The large contiguous sky coverage and sheer number

---

[1] Note that these reference lists are meant only to give some representative Galactic (sub-)structure studies from SDSS, and are far from complete.

of targets that will be observed by LAMOST can help to overcome the limitations of SEGUE for studies of Galactic structure; however, this requires that the selection of LEGUE targets be done in a well-understood, simply-defined manner.

Other spectroscopic surveys of large numbers of stars have focused on magnitude-limited samples. For example, the Radial Velocity Experiment (RAVE; Steinmetz et al. 2006), is a survey of $\sim$ one million stars to a limiting magnitude of $I = 12$ in the southern hemisphere. Upcoming surveys, such as the HERMES Galactic Archaeology project (e.g., Barden et al. 2008; Freeman & Bland-Hawthorn 2008; Freeman 2010) will also observe magnitude-limited samples of stars (in this case, to $V = 14$). Obviously, a magnitude limited survey does not require careful selection of subsets of available targets, as is required for deeper surveys such as SEGUE or LAMOST.

In this paper, we present a general target selection algorithm designed for surveys such as LAMOST where the number of available candidates is much larger than the number of objects to be observed. The method is sufficiently general to be extensible to any target selection process, and can use any number of observables (i.e., photometry, astrometry, etc.) to perform the selections. The paper is organized as follows: we introduce a general target selection algorithm and show some examples of different selection biases that can be applied. We follow this with a hypothetical survey design, discussing the priorities for target selection in this mock survey, then show examples of the adopted target selection parameters for a moderate latitude ($b \sim 30°$) and a high latitude ($b \sim 60°$) field. This hypothetical survey has target priorities similar to those outlined by Deng et al. (2012), based on LEGUE's science goals. More details of the use of our target selection algorithm for the LEGUE pilot survey can be found in Yang et al. (2012), which discusses the dark nights portion of the pilot survey, and Zhang et al. (2012), where a summary of the bright nights observing program is given (see also Chen et al. 2012 for discussion of an alternative target selection process that was applied to the Galactic disk portion of the LEGUE pilot survey). We follow the example survey illustration with some discussion about the difficulty in recreating a "statistical sample" of stellar populations from the observed set of spectra. The algorithms developed here have been used mostly with Sloan Digital Sky Survey (SDSS) photometry as inputs (though see Zhang et al. 2012 for an example using 2MASS data), but in practice any photometric, astrometric, spectroscopic, or other data known about the input catalog stars can be used in the selection process. The target selection programs discussed in this work were developed in the programming language IDL.

2 TARGET SELECTION ALGORITHMS

We initially set out to solve the general spectroscopic survey target selection problem: starting with an input catalog of stars with any number of "observables" (for example SDSS, with *ugriz* magnitudes, positions, proper motions, etc.), define a general target selection algorithm that is capable of producing the desired distribution of targets. The assumption is that one begins with a large input catalog, where the number of sources is larger than the number of objects that can be observed with LAMOST. An input catalog would be a data table with N_S stars for which we have N_O observables (such as right ascension, declination, magnitude, color, proper motion component, etc.)

$$\lambda = [\lambda_i]_j, \tag{1}$$

where $j = 1, 2, ..., N_S$ denotes any one of N_S stars in the input catalog, and $i = 1, 2, ..., N_O$ denotes any of the N_O observables which are available for every star. To select targets for a spectroscopic survey such as LEGUE, one would minimally require sky coordinates and a magnitude ($N_O \geq 3$).

For every LAMOST field, a number of stars can be randomly selected as targets (based on how many fibers are available) among stars which are located in the field, and for which each was assigned a statistical weight. This statistical weight can be assigned according to a function $P = P(\lambda_1, \lambda_2, ..., \lambda_{N_O})$ of the N_O observables, such that the probability for selecting star j as a target can be expressed as

$$P_j = P([\lambda_1]_j, [\lambda_2]_j, ..., [\lambda_{N_O}]_j) \tag{2}$$

with the requirement

$$\sum_j^{N_S} P_j = 1. \tag{3}$$

The trivial case would be for every star to have the same probability of being selected ($P_1 = P_2 = ... = P_j, \forall\, j$). Calling this trivial case "model A," and denoting its probability function $P_{j,A}$, we then have

$$P_{j,A} = (N_S)^{-1}. \tag{4}$$

Alternatively, one could base the selection on the statistical distribution, Ψ_0, of the values taken by the observables. Defining Ψ_0 as a continuous function over the N_O observables, one would have

$$\Psi_0 = \Psi_0(\lambda_1, \lambda_2, ..., \lambda_{N_O}) \equiv \Psi_0(\lambda_i), \tag{5}$$

which represents the density of recorded values for the observables, normalized following

$$\int \Psi_0(\lambda_i) \prod_i^{N_O} d\lambda_i = 1. \tag{6}$$

One way to determine Ψ_0 for the input catalog would be to calculate the local density of sources at $(\lambda_1, \lambda_2, ..., \lambda_i)$, estimated by counting the number of stars j whose observables satisfy the condition

$$\sqrt{\sum_i (\lambda_i - [\lambda_i]_j)^2} < \Delta\lambda, \tag{7}$$

where $\Delta\lambda$ defines the size of the volume in the space of observables over which the stars are being counted, i.e., the resolution of the function Ψ_0. For example, one could determine the density function $\Psi_0 = \Psi_0(g, g - r)$, calculating how many stars can be found within 0.1 magnitudes of the parameter space location $(g, g - r)$, which would mean using $\Delta(g, g - r) = 0.1$ mag. This can be extended to any number of observables to define a "density" over multiple parameters; an example would be using additional colors, calculating the number of stars within 0.1 magnitude of $(g, u - g, g - r, r - i,...)$.

An example of a Ψ_0 function can be seen in Figure 1, which shows the statistical distribution of $g - r$ color for all stars in two sample LAMOST plates as the solid black histograms. These have been normalized so that the sum of all bins equals one, and can thus be thought of as probability functions (in this case, $\Psi_0 = \Psi_0(g - r)$). A similar plot is seen in Figure 2 for r magnitudes in the same two plates. For the remainder of this paper, we will use examples from the simple case of the local density defined by the number of stars within 0.1 magnitudes in r, $g - r$ and $r - i$, i.e. $\Psi_0 = \Psi_0(r, g - r, r - i)$.

It is useful to point out that if one performs a target selection following method "A," then the density distribution of the target stars, denoted Ψ_A, would be approximately the same as the statistical distribution in the input catalog, to within the Poisson errors, i.e. $\Psi_A(\lambda_i) \approx \Psi_0(\lambda_i)$.

Typically, however, one may want to obtain a list of targets whose statistical distribution differs from the distribution in the input catalog. For instance, one may want to overselect objects in a given color/magnitude range, or pay more attention to outliers or unusual stars. One possibility, which we will call "Method B," is to assign a selection probability that is inversely proportional to the local density in the space of observables

$$P_{j,B} = \frac{K_B}{\Psi_0([\lambda_1]_j, [\lambda_2]_j, ..., [\lambda_{N_O}]_j)}, \tag{8}$$

where K_B is a normalization constant to ensure that $\sum_j P_{j,B} = 1$. In Method B, the statistical distribution of the selected stars (Ψ_B) over the observables (λ_i) is different from that of the input
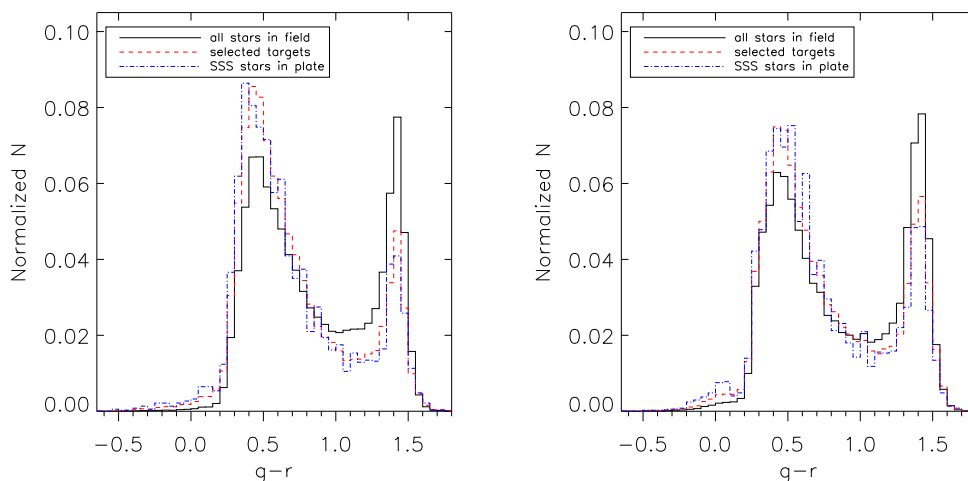
Fig. 1 Fractional distribution of $g - r$ color for stars in two sample regions of the sky $6° \times 6°$ in size (slightly larger than the area of a LAMOST plate); all data are selected from SDSS DR8. The left panel is for a field spanning $(\alpha, \delta) = (130° - 136°, 0° - 6°)$, corresponding to a field center in Galactic coordinates of $(l, b) \approx (225°, 28°)$. This field contains a total of 102 199 stars between $14.0 < r < 19.5$. The right panel shows a field at $(\alpha, \delta) = (170° - 176°, 0° - 6°)$, or $(l, b) \approx (261°, 59°)$, containing a total of 44 566 stars. In both panels, the black solid line represents all stars in the field. The dashed histogram represents the stars selected by our algorithm as input to the LAMOST fiber-assignment program ($\sim$600 per square degree), and the dash-dotted line shows the resulting distribution of spectroscopic targets in a single LAMOST plate ($\sim$ 4000 stars assigned to fibers). Each histogram has been normalized to one, so that the bin heights represent the fraction of targets within each bin.

catalog (Ψ_0), and in fact it is to first order uniform over all values of λ_i, i.e. $\Psi_{\rm B}(\lambda_i) \approx K_{\rm B}$. Examples of this type of selection are seen in panels (b) of Figures 3 and 4. Note, however, that because the local density was calculated using r, $g - r$ and $r - i$, the distribution does not look uniform in the color-magnitude diagrams (top and middle rows). However, the distribution in three-dimensional parameter space defined by $\lambda_i = r, g - r, r - i$ should be roughly uniform.

As a generalization, one can assign probability that is inversely proportional to some power of the local density, i.e., $[\Psi_0]^{-\alpha}$. In this case, which we will call Method C,

$$P_{j,\rm C} = \frac{K_{\rm C}}{[\Psi_0([\lambda_1]_j, [\lambda_2]_j, ..., [\lambda_{N_{\rm O}}]_j)]^\alpha} \tag{9}$$

where $K_{\rm C}$ is a normalization constant to ensure that $\sum_j P_{j,\rm C} = 1$. One can now see that Methods A and B represent special cases where $\alpha = 0$ and $\alpha = 1$, respectively. The random selection approach ($\alpha = 0$) would be ideal if one simply wanted a selection that samples all of parameter space with the same frequency as the input catalog. A weighting by $1/\Psi_0$ ($\alpha = 1$), on the other hand, produces an output catalog that samples the space of observables more evenly, and thus contains a much larger fraction of "rare" objects (i.e., those in less-populated regions of parameter space, and de-emphasizes regions of higher local density relative to the input catalog; see, e.g., panels (b) of Figs. 3 and 4). Adopting a value $0 < \alpha < 1$ would result in a selection intermediate between these two scenarios; one that increases the chances of rare objects entering the selection, but still robustly samples the high-density regions of the input distribution. The effects of using $\alpha = 1/2$ are seen in panels (c) of
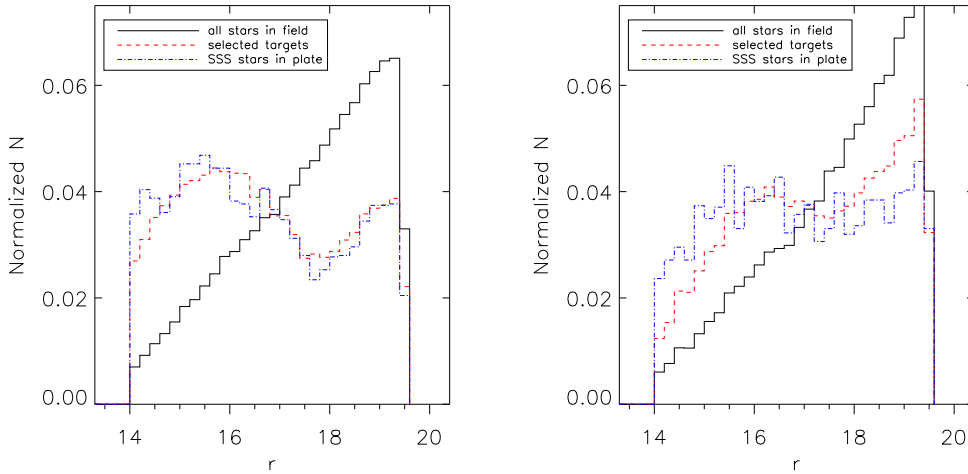
Fig. 2 Fractional distribution of r magnitude for stars in the same two example regions as in Fig. 1. The line styles and colors are also the same as in Fig. 1. Each histogram has been normalized so that the bin heights represent the fraction of targets within each bin; this distribution can be thought of as a probability distribution, $\Psi_0(r)$, of finding a star in each magnitude range.

Figures 3 and 4. In the case where one would like to place a particular emphasis on rare objects, a value of $\alpha > 1$ could be used, which would introduce a bias against the selection of more common objects as targets. (Also, note that if there are fewer stars in a given region of parameter space than are selected in the more densely populated regions, that region will continue to be under-dense no matter what emphasis is applied.)

In any case, one may also be interested in over-selecting objects occupying a particular region of parameter space (for example, a narrow color or magnitude range, or simply selection of more blue than red stars). To achieve this, an overemphasis can be included that favors the selection of stars in a particular range of an observable by including a bias explicitly in the selection probability function. Any functional form of each of the observables, λ_i, can be introduced to achieve the desired effect

$$P_{j,\mathrm{D}} = \frac{\prod_i f_i([\lambda_i]_j)}{[\Psi_0([\lambda_1]_j, [\lambda_2]_j, ..., [\lambda_{N_\mathrm{O}}]_j)]^\alpha}. \tag{10}$$

Here the $f_i(\lambda_i)$ can be any function of the observables λ_i.

Two examples that are currently implemented for the LAMOST pilot survey are a local emphasis over a specific range of colors, and a general bias over the magnitude range to emphasize brighter stars or fainter stars. A local emphasis is achieved using a function of the form

$$f_i(\lambda_i) = 1 + A_i \mathrm{e}^{-\frac{(\lambda_i - x_i)^2}{\sigma_i^2}}, \tag{11}$$

where x_i is the central value of interest for the observable λ_i (to emphasize the center of the color or magnitude range), σ_i is the range of interest, and A_i is the "over-selection" factor (how strongly you wish to overemphasize these objects compared to stars outside of this range). An example is shown in panels (d) of Figures 3 and 4, where the region of interest is centered at $g - r = 0.8$, with a range $\sigma_{g-r} = 0.2$ and an overemphasis factor $A = 10$.
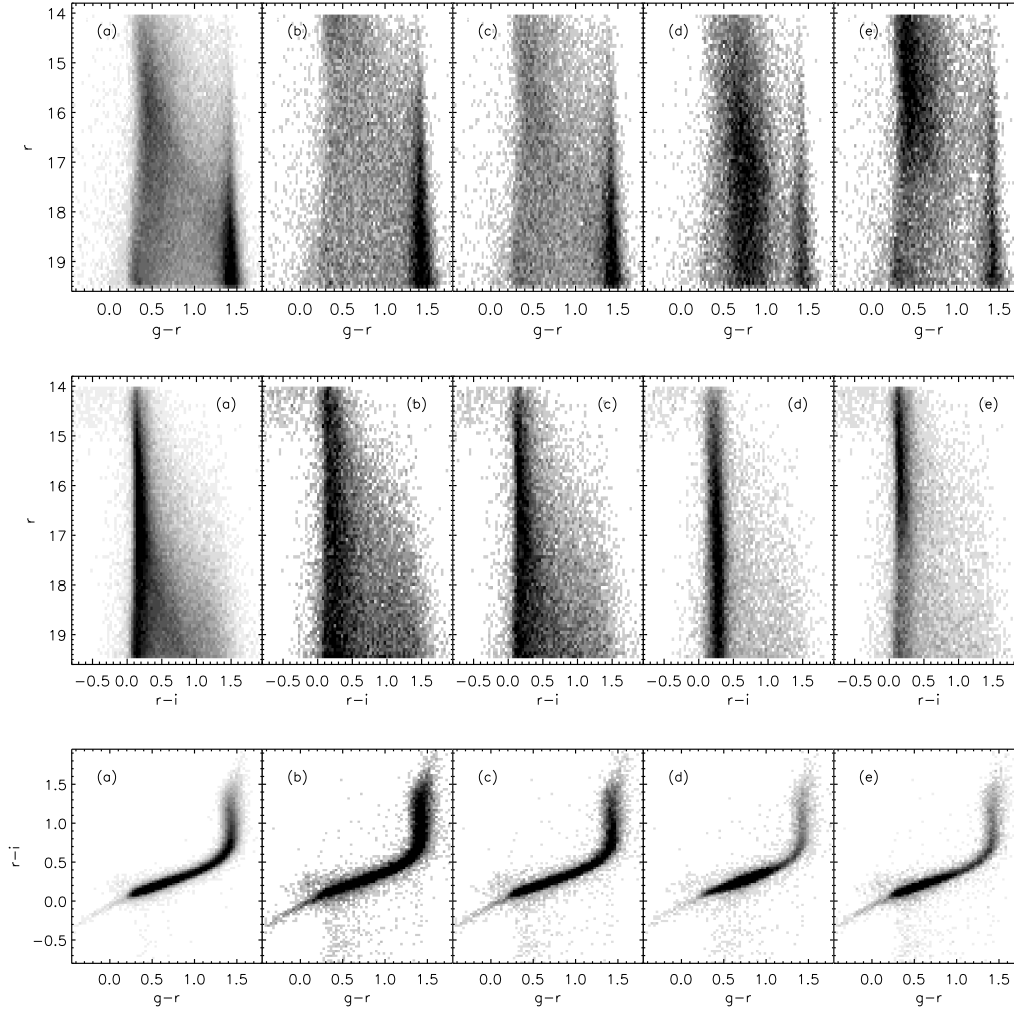
Fig. 3 Color-magnitude Hess diagrams for different selections of stars in the region $(\alpha, \delta) = (130° - 136°, 0° - 6°)$, corresponding to a field center in Galactic coordinates of $(l, b) \approx (225°, 28°)$. Panel (a) shows all of the 102 199 SDSS stars between $14.0 < r < 19.5$ in this field of view. Panels (b)–(e) show the results of selecting 600 stars per square degree in this field, with different selections (based on local density in three-dimensional $r, g−r, r−i$ space, or $\Psi_0(r, g−r, r−i)$) represented in each panel. Panel (b) stars were selected with $\alpha = 1$, and panel (c) depicts the $\alpha = 0.5$ case. The $\alpha = 1$ selection (i.e., weighting by the inverse of Ψ_0) in panel (b) strongly de-emphasizes high-density regions of the CMD in favor of rare stars. M-stars at $g−r \sim 1.5$ appear oversampled in this figure; this arises because they are more spread out in $r − i$ colors than in $g − r$, causing them to be emphasized by the density weighting in 3-D parameter space. The overemphasis of rare objects is slightly less pronounced for $\alpha = 0.5$ (panel (c)), with a significant number of stars selected from the high-density regions of the CMD. In panel (d) we illustrate the results of selection with $\alpha = 0.5$ and with an over-selection of the region centered at $g − r = 0.8$ with width $\sigma_{g−r} = 0.2$ and overemphasis factor $A = 10$. This selection produces an overselection of stars centered at $g − r = 0.8$, while retaining some stars from the remaining parameter space. Finally, panel (e) illustrates a selection with $\alpha = 0.5$ and a linear bias in color beginning at $g − r = 1.1$ and sloping upward toward bluer colors with slope 2.5, and also a linear magnitude emphasis beginning at $r = 17.5$ with slope 1.0 toward brighter magnitudes.
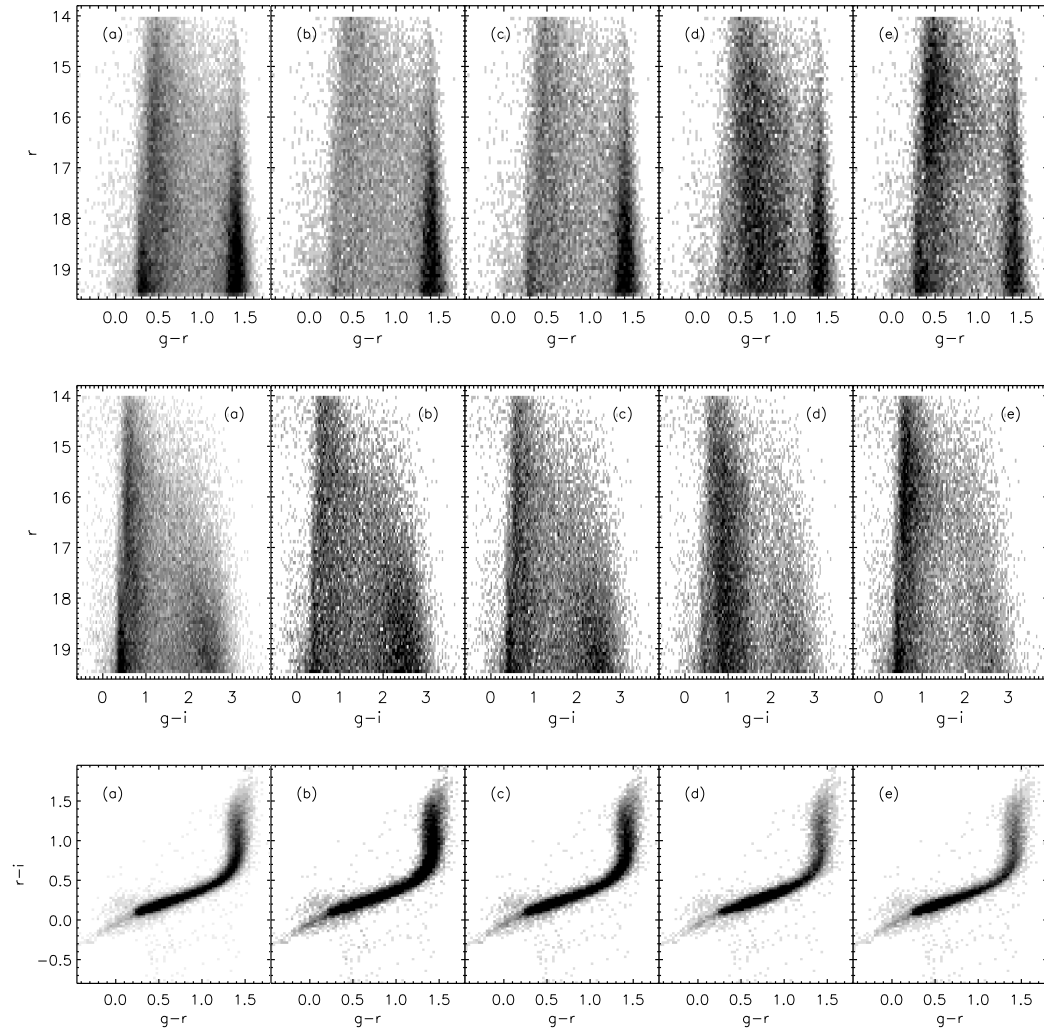
Fig. 4 As in Fig. 3, but for a different field located at $(\alpha, \delta) = (170^\circ - 176^\circ, 0^\circ - 6^\circ)$, or $(l, b) \approx (261^\circ, 59^\circ)$. This higher-latitude field contains a total of 44 566 SDSS stars between $14.0 < r < 19.5$.

Likewise, a general bias is introduced by the use of a linear function of the form

$$f_i(\lambda_i) = 1 - m_i(X_i - \lambda_i), \tag{12}$$

where m_i is the slope of the linear emphasis function, and X_i is the limiting value where the linear emphasis ends (either the minimum or maximum allowed value of λ_i). This produces a function that is 1.0 at one extremity, and increases to higher values from the limiting value X_i. An example of this would be to use a linear function that increases toward lower values to overselect stars of bluer colors and/or brighter magnitudes. The effects from this type of selection bias are shown in panels (e) of Figures 3 and 4; the color selection is anchored at $g - r = 1.1$ with slope 2.5 increasing toward bluer colors, and the magnitude emphasis has slope 1.0 anchored at $r = 17.5$, increasing toward the bright end.

2.1 Selecting Stars Using the Assignment Probabilities

Here we describe the method we have implemented to select target stars based on the selection probability function (i.e., P_j). Once the assignment probability for each star in the input catalog has been defined, a cumulative probability is calculated for each star in the list which consists of the sum of the probabilities of all stars in the list up to (and including) that particular star

$$P_{\mathrm{cum},j} = \sum_{k=0}^{j} P_{\mathrm{cum},k},$$

(13)

where "j" is the index of each star. A random number is then generated (using a uniform distribution from 0.0 to 1.0), and one star from the list is identified for which the random number is less than the cumulative probability, but greater than the cumulative probability of the previous element (where $P_{\mathrm{cum},j-1} < \mathrm{randnum} < P_{\mathrm{cum},j}$). This star is placed in the list of "selected" stars and removed from the sorted list of candidates for selection. The assignment probabilities are then renormalized so that they sum to unity once more, and the cumulative probabilities are recalculated. (Note that the order the stars appear in the input catalog does not matter, since a star with a larger selection probability will carve out a larger range of the cumulative probability space, and thus be more likely to be selected.) The process is repeated until the desired number of stars has been selected. Selecting targets in this manner has the effect of preferentially choosing stars with higher selection probabilities, but still selecting some stars from the entire range of parameter space. This also means that the stars selected near the beginning will have different "demographics" (i.e., occupy different distributions in parameter space) than those selected later. Thus if a region of sky is revisited for a second (or more) observation, the distribution of targets in the observables will differ from the overall distribution in that same field of view. This in turn means that the selection probability as a function of the observables will also differ when revisiting a region of sky.

In the case of the LAMOST pilot survey, the number of targets to select is set by the fiber-assignment software's requirement that the input catalog contain roughly three times the desired number of spectroscopic targets. Thus, since LAMOST contains 200 fibers per square degree in the focal plane, we select 600 stars deg^{-2} in the input catalogs (though this target density is a parameter that can be set when running the program). This is achieved by dividing the sky into $2° \times 2°$ blocks, and selecting targets in each block until the target density has been reached. Defining the local density separately for each of these blocks has the benefit of mitigating the effects of large-scale spatial variations of stellar populations within the survey footprint on the defined local densities (and thus the target selection probabilities). Note, however, that because three times the fiber density is required in the input catalog, only 1/3 of the targets selected for input to the fiber assignment algorithm will be observed. If the target selection was being done at the same time the fibers were being assigned to objects, one could maximize the probability that objects in a particular parameter range were selected by assigning a very large probability in that range. Because the probabilities are pre-assigned separately from the fiber assignment process, we have implemented a priority scheme to preserve information about which objects would have been selected first. In the absence of this priority scheme, all objects sent to the fiber assignment algorithm would be observed with a probability of $\sim 1/3$, so it would be impossible to regularly observe more than 1/3 of any type of object. Of course, revisiting the same plate multiple times increases the chance of observing all objects with certain selection criteria, since those that were unable to be assigned on the first plate can be picked up on later observations.

The LAMOST target assignment algorithm allows us to assign priorities from 0–99 for each of the selected objects, with lower numbers indicating higher priority for selection by a fiber. The probability for selection calculated by the target selection algorithm must be converted to an integer priority value for the fiber assignment program, rather than being used directly to assign targets to fibers. When each fiber is being assigned a target, all of the possible targets within its patrol

radius are examined, and the one with the lowest priority value is assigned. If all targets have been given equal priority, then the fiber will be assigned to the target closest to its "home" position; such an instance would thus produce a uniform spatial distribution of targets, ignoring any selection preferences based on photometry or other properties. This also means that if multiple high-priority targets are within the patrol radius of a single fiber, only one of them will be assigned to a fiber. To ensure that the desired target distribution in parameter space is achieved, one would ideally assign as many priority values as possible, so that the probability distribution created by the target selection algorithm would be closely followed. To do this, we assign priorities from $1-M$ (where $M < 99$; for the pilot survey we used $M = 80$, with the remaining priorities being reserved for other possible uses) to the 600 stars selected in each square degree of sky. The priorities are assigned by taking the total number of targets desired in a given region (in this case, 600 per square degree), dividing by M, then looping from $1-M$, and assigning this number of targets to each priority value. Since the probability weighting should preferentially select targets of higher interest at the beginning of the selection process, this method ensures that lower priority values (i.e., higher chance of being assigned) are predominantly given to the objects with high probability for assignment, with lower-probability stars mostly having priorities that will make them less likely to be assigned. In practice, this complicates the statistical understanding of the target distribution, but our tests have shown that this method in effect reproduces our desired target distributions.

3 A SAMPLE HYPOTHETICAL SURVEY

3.1 Survey Goals and Target Categories

With a general target selection algorithm developed, we now explore the question of what combinations of parameters can be used to achieve different target selection goals. In a non-magnitude-limited spectroscopic survey with multiple science goals (for example, LEGUE), there will typically be certain types of objects that are valued more than others. An example might be blue horizontal-branch (BHB) star candidates. These are an extremely valuable resource for Galactic structure studies because they are relatively rare, intrinsically bright (making them ideal probes of the distant Milky Way halo), trace metal-poor populations typical of the halo, easy to derive distances for, and occupy regions in photometric colors that are not confused with many other types of objects. So, for example, a study that is interested in BHB stars could try to simply select all stars with SDSS colors $(g - r) < 0.0$ and $(u - g)$ colors unlike those of QSOs. With our target selection algorithm, these objects could easily be preferentially selected without resorting to something like an abrupt cutoff at a certain photometric color. This can be achieved in one of two ways (or a combination of both): first, an emphasis on rare objects (BHB stars have relatively low densities in color-magnitude or color-color diagrams; see, e.g., Fig. 1 for an illustration of the paucity of such blue stars) can be achieved via weighting by the inverse of the local density in color/color/magnitude space (or, even better, weighting by $\alpha = 1/2$), and secondly, by adding a linear emphasis that increases blueward of some cutoff color. Examples of the density weighting are shown in panels (b) and (c) of Figures 3 and 4, illustrating the effect of weighting by $\alpha = 1$ and $\alpha = 1/2$, respectively. The number of rare, blue objects is enhanced in these relative to the fraction of blue stars in the input catalog. An additional linear weighting can be applied; for example, one could choose to multiply the local density by a linear function beginning at $g - r = 0.3$ and increasing blueward. Both of these methods will increase the number of blue stars selected, while avoiding an abrupt cutoff for the selection at $g - r = 0.0$.

Within a given collaboration, there may be many science goals (for example see Deng et al. 2012 for a discussion of LEGUE's science aims). Balancing the need for a well-understood selection function with numerous target types is easily done with the method we have outlined. Here we create a hypothetical survey to use as an example. Our example survey (which happens to very closely resemble many of LEGUE's goals) aims to study the Galactic halo, while also sampling a large

number of nearby stars of all types for studies of the Galactic disk. The survey will use only SDSS photometry for target selection, with no constraints on proper motions or other properties, selecting among stars with $14 < r < 19.5$.

Studying the halo requires intrinsically bright, easily identified tracers such as BHB stars or K/M-giants to probe to large distances, and also a large sample of F-type turnoff stars at all magnitudes. F-type stars occupy a color range of relatively unambiguous luminosity classification (other than the occasional asymptotic giant-branch star). The BHB stars are very blue ($g - r < 0.0$), while K/M giants are red stars with $g - r > 1.0$. F-turnoff stars have colors of roughly $0.2 < g - r < 0.5$. BHB stars (and to a lesser extent, K/M giants) occupy a region of low stellar density in color-magnitude space, while F-turnoff stars are very common. This proposed survey would also like to obtain spectra of metal-poor M subdwarfs, which are not distinguishable in the $g - r$ band from all other M-dwarfs, but separate clearly in $r - i$ colors, while still sampling a large number of nearby M-dwarfs that can be used to probe local kinematics. There is also interest in following up interesting discoveries with high-resolution spectroscopy, so we wish to overemphasize bright stars within reach of echelle-resolution spectrographs. Finally, there is a desire to sample all stellar populations, but preferentially observe rare objects first, in order to open up the discovery space to rare (and perhaps previously unknown) stellar types. Briefly, then, the target selection categories are as follows:

– Sample a larger fraction of "rare" stars than the "less rare" cases. As shown in Section 2, this is exactly what is achieved by the use of local density weighting, $P_j \propto [\Psi_0(\lambda_i)^{-\alpha}]_j$, in assigning selection probabilities to each star. In particular, we have shown examples of $\alpha = 1$ and $\alpha = 1/2$; the $\alpha = 1$ case predominantly selects rarer objects (i.e., it under-emphasizes parameter spaces of high stellar density) than the $\alpha = 1/2$ density weighting.
– Select nearly all stars with $0.1 < (g - r) < 1.0$ and $r < 17$ at high Galactic latitudes, and sub-sample at $b < 40°$.
– Select nearly all stars with $g - r < 0.0$ and $u - g$ colors that are unlike those of quasars (BHB and blue straggler candidates).
– Select a significant fraction of the stars with $0.0 < (g - r) < 1.0$, $17 < r < 19.5$ and $u - g$ colors that suggest they are not quasars. The bluer side of the color range should be selected with a probability about twice the redder side of the range to emphasize the F-type turnoff stars.
– Select a large number of M dwarfs at all magnitudes.

We note that the above discussion refers to a survey with multiple visits to each sky position, which can thus meet the requirements of nearly-complete samples of some subsets of stellar types (for example, the very blue $g - r < 0$ stars). However, in practice not all of the high priority targets can be placed on one observation due to constraints on fiber positioning. Thus, for a survey where only a single visit to each sky area is planned, these "requirements" should be considered to mean that one would like as many as possible of the stars in these categories.

3.2 Adopted Target Selection Parameters

Through many tests, it was determined that the simplest combination of parameters striking a good balance between all these desired categories of targets is

– $\alpha = 0.5$, which weighs by the inverse square root of the local density.
– Linear ramp bias function in $g - r$, beginning at $g - r = 1.1$ and increasing blueward with a slope of 2.5.
– Linear ramp bias function in r magnitude, beginning at $r = 17.5$, increasing toward brighter stars with slope of 1.0.

The particular values of these parameters (and particularly of $\alpha = 0.5$) were determined somewhat subjectively based on visual examination of the selected targets and statistical distributions of targets
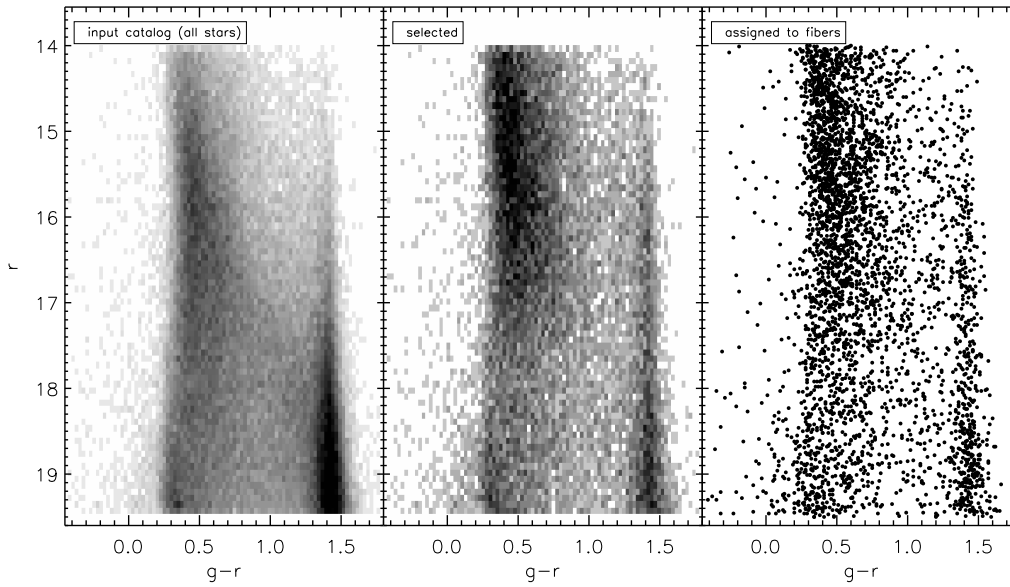
Fig. 5 *Left panel:* Color-magnitude Hess diagram of all 112 099 stars between $14 < r < 19.5$ selected from SDSS DR8 at $(\alpha, \delta)_{\mathrm{J2000}} = (130° − 136°, 0° − 6°)$. *Center panel:* Result of selecting 600 targets per square degree from the stars in the left panel. The target selection used $\alpha = 0.5$ (to emphasize rare objects) with a ramp in color (to increase the fraction of blue stars selected) anchored at $g − r = 1.1$ and increasing blueward with slope of 2.5, and a ramp in magnitude (to weight bright stars more heavily) starting at $r = 17.5$ and increasing with slope of 1.0 toward the bright end. This selection represents 21.1% of the stars in the field of view. *Right panel:* Distribution of targets assigned to LAMOST fibers upon running the catalog from the center panel through the fiber-assignment software. This panel contains a total of 3715 stars, or 3.6% of the total number within the field of view.

from separate categories (as seen in Table 1, which will be discussed in more detail below). Of course, detailed analysis could be done to optimize these target selection parameters if desired. However, in the case of a survey such as LAMOST, which will observe large numbers of stars with a variety of science goals, we simply select these parameters to produce input catalogs that are broadly consistent with the desired target distributions and sample all of parameter space to some extent.

Table 1 Fraction of Stars from Each Target Category Assigned to Fibers in a Single LAMOST Plate

RA (°)	Dec (°)	l (°)	b (°)	Total stars	Assigned	% Assigned (%)	Very blue (%)	Bluish, Bright (%)	Bluish, Faint (%)	Red (%)
133±3	3±3	225	28	102199	3715	3.6	22.6	6.7	2.7	2.1
173±3	3±3	261	59	44566	3722	8.4	27.0	15.5	6.9	5.7

3.3 Sample Target Selections

In this section, we show examples of outputs from the target selection code, and follow this by selecting stars for targeting from among these using the LAMOST fiber-assignment routine. These
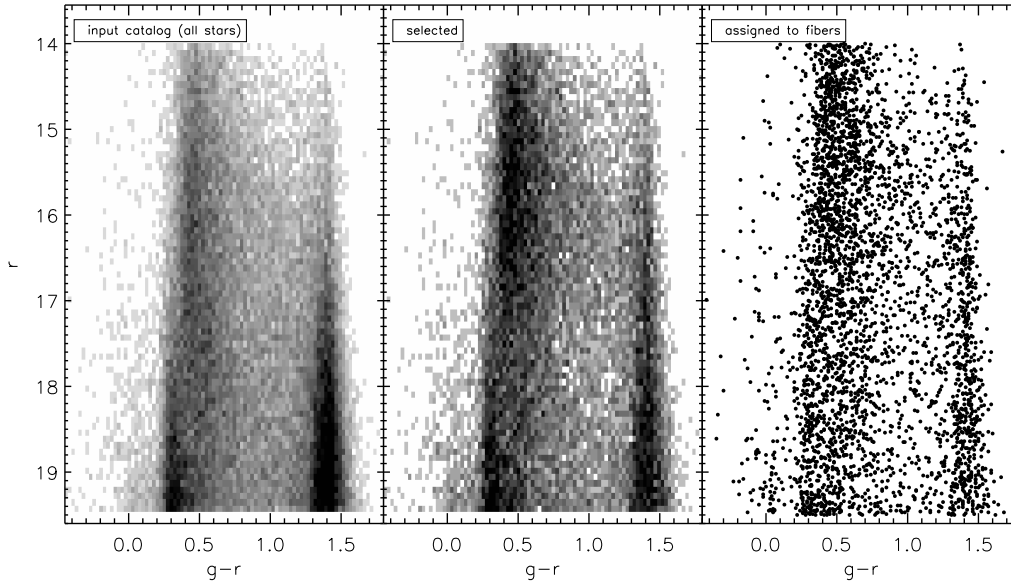
Fig. 6 As in Fig. 5, but for the higher-latitude field at $(\alpha, \delta)_{J2000} = (170° - 176°, 0° - 6°)$. This field of view contains a total of 44 566 stars, of which 48.4% were selected for the center panel. Of these, 3722 (seen in the right panel), or 8.4%, were assigned to fibers.

examples use the data from the same two fields presented earlier in this work, selected from $6 \times 6°$ fields at $(\alpha, \delta)_{J2000} = (130° - 136°, 0° - 6°)$ and $(\alpha, \delta)_{J2000} = (170° - 176°, 0° - 6°)$. These fields were chosen to show an example of a field at somewhat low latitude ($b \sim 30°$), and another at high latitude ($b \sim 60°$). The local density, Ψ_0, is calculated in each of these fields using r magnitudes and $g - r$, $r - i$ colors. For reference, the r vs. $g - r$ color-magnitude distribution of all 102 199 stars in the $\alpha \sim 133°$ field of view is seen in the left panel of Figure 5, and the 44 566 stars in the lower-latitude $\alpha \sim 173°$ field in Figure 6 (note that these are the same as panels (a) in Figs. 3 and 4).

The center panels of Figures 5 and 6 show the results of running the target selection routine on the input catalogs, using $\alpha = 1/2$, a linear ramp in color, beginning at $g - r = 1.1$ and rising blueward with slope 2.5, and a ramp in magnitude, increasing from $r = 17.5$ with slope 1.0 toward bright stars. These selections contain 600 stars per square degree, the required target density to be input into the fiber assignment program. In the field at $\alpha \sim 133°$ (Fig. 5), 21.1% of the 112 099 total stars in the region were selected as candidates, and in the higher-latitude $\alpha \sim 173°$ field (Fig. 6) this number rises to 48.4% of the total available stars.

After running the catalogs selected for these two fields of view through the LAMOST fiber assignment program, ~ 3700 stars in each of the two fields are allocated to fibers (the remaining fibers are to be used for sky and other calibration purposes). The right panels of Figures 5 and 6 show the stars assigned to fibers in these two fields. Generally, it is clear that quite a few "rare" objects (for example, at intermediate colors of $0.8 < g - r < 1.2$, or bright M-star candidates at $g - r \sim 1.4$) are selected by this method, but that densely-populated regions of color-magnitude space are also well-sampled. Note the fairly dramatic overemphasis of bright ($r < 17$), blue ($g - r < 1.0$) stars achieved by the weighting scheme.

To assess how well the algorithm achieved the list of target selection goals outlined in Section 3.1, we select stars from the color and magnitude ranges in which specific target-selection

goals were focused, and explore the relative emphasis or de-emphasis achieved by our code. The degree of emphasis can be seen by comparing the fraction of stars selected within a given target category to the fraction of the total number of stars in the field. These results are given for the two example fields in Table 1. The table lists the number of stars in each of the two fields of view ("total stars"), followed by the number assigned to fibers for a single LAMOST plate ("assigned"), and the percentage of the total stars that were assigned to be observed ("% assigned," or "assigned"/"total stars"). The following four columns represent the four categories outlined in Section 3.1: "very blue" stars with $g - r < 0$, "bluish, bright" stars with $0.0 < g - r < 1.0$ and $r < 17$, "bluish, faint" stars with $0.0 < g - r < 1.0$ and $r > 17$, and "red" stars with $g - r > 1.0$. In each of these columns, we provide the percentage of the total number of stars in the field that satisfy those criteria that were assigned to a fiber on the plate. This percentage can be compared to the "% assigned" column to see over- or under-emphasis; i.e., if the target selection was uniform across color-magnitude space, one would expect roughly the same fraction of stars to have been assigned in each category. Thus, for the $\alpha = 133°$ field, the fact that 22.6% of the very blue stars were assigned compared to 3.6% overall means that the "very blue" stars have been overemphasized by a factor of > 6. Examination of Table 1 shows that, at least broadly, we have achieved our goals of strongly over-selecting very blue objects, increasing the fraction of bright, blue stars that gets observed, yet still retaining a significant number of faint, bluish stars and red K- and M-star candidates (note that > 800 red stars with $g - r > 1.0$ were assigned in each plate – even though they have been underemphasized, they are still well-represented).

4 SOME CAVEATS ABOUT STATISTICAL TARGET SELECTION

Ostensibly, one of the reasons for having a smoothly-varying, well understood selection function is to be able to infer the underlying stellar populations from a given set of spectroscopically observed stars. However, in order for this to be possible, detailed records of the entire target selection and fiber assignment process need to be kept. The first issue affecting this is the need to supply the fiber assignment routine with a catalog with higher target density than the fiber density on the sky. Because of this, not all high priority targets will be placed on fibers. Some fibers in each observation will inevitably fail to yield useful spectra, making it necessary to factor the "missed targets" into the analysis.

Of course, any routine that preferentially targets certain objects will produce different target demographics if multiple visits to the same sky position are desired. An illustration of this is seen in Figure 7, which shows examples of three LAMOST plates selected in each of the two example fields used throughout this work. The upper panels show the relatively low-latitude ($b \sim 30°$) field, and the lower panels show the $b \sim 60°$ field. The target selection parameters were the same as those used in Sections 3.2 and 3.3, and seen in Figures 5 and 6. In each row, the five panels show r vs. $g - r$ Hess diagrams of (a) the input sky distribution from SDSS, (b) the first plate selected, (c) the second plate selected (excluding stars from plate 1), (d) the third plate selected (excluding stars from plates 1 and 2), and (e) the sum of all three plates from (b)–(d). In the low-latitude field there are many stars available, so that the effects of the preferential targeting categories are seen in all three plates (especially the emphasis on bright, blue stars). However, the sum of all three plates (panel (e)) contains representative samples from all regions of color-magnitude space. The higher-latitude field (lower panels) is quite different. The stellar density is much lower in this field, so that in three LAMOST plates, a total of 24.4% of the stars between $14 < r < 19.5$ are assigned to fibers. The first plate (panel (b)) appears very similar to the corresponding selection from the low-latitude field, with bright, blue stars overemphasized (note also that many of the very blue, $g - r < 0.2$ objects are gone after the first plate). By the second and third plates in this field, however, a large fraction of the bright, blue stars have already been assigned, and the selected stars start to cover more of the parameter space (specifically, there are many more faint stars – especially a lot more faint, red
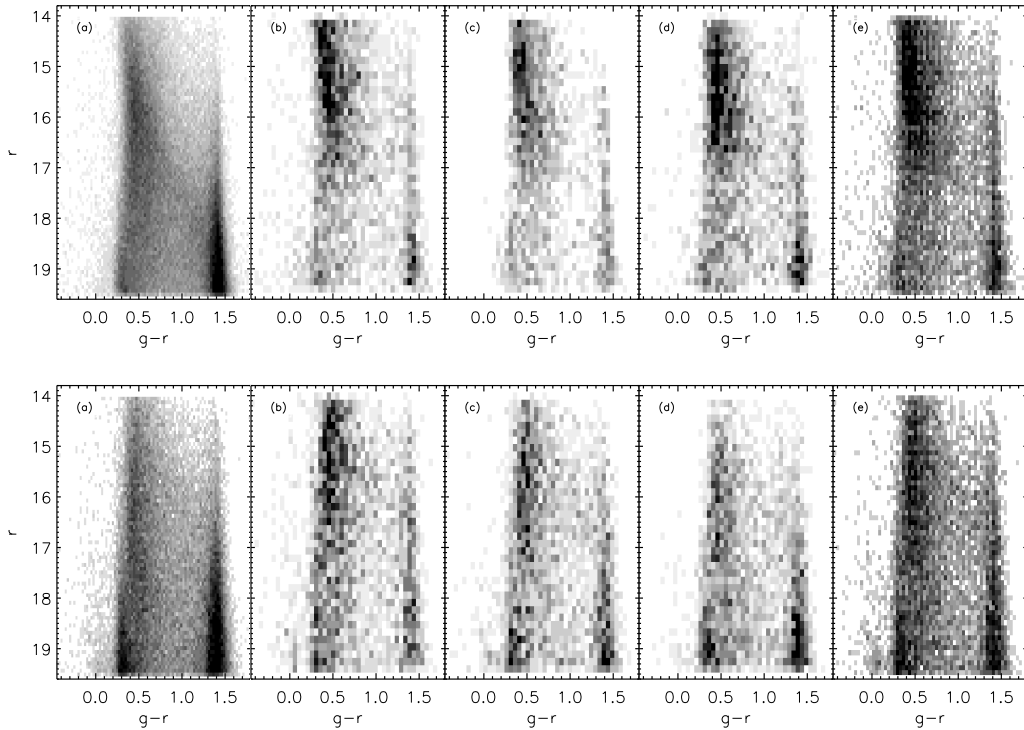
Fig. 7 Color-magnitude Hess diagrams for stars selected from the low-latitude ($b \sim 30°$) field (upper panels) and high-latitude ($b \sim 60°$) example fields. The same target selection parameters were used as in Figures 5 and 6. In each row, the panels represent (a) all stars from SDSS in the field of view, (b) stars selected for fiber assignment on the first LAMOST plate in this field, (c) a second plate excluding the stars in the first assignment, (d) a third plate, excluding stars from the first two, and (e) the sum of the three selected plates. The high-latitude field of view contains a total of 44 566 stars, and the lower-latitude field has 102 199. Roughly 3700 stars were assigned on each of the three plates, so that in total, 24.4% of the high-latitude stars were assigned to a fiber on one of the three plates, and 10.8% of those at lower latitudes. In the low-latitude field there are many stars available, so the effects (especially the emphasis on bright, blue stars) of the preferential targeting categories are obvious in all three plates. However, the sum of all three plates (*upper panel* (e)) contains representative samples from all regions of color-magnitude space. The higher-latitude field (*lower panels*) has much lower stellar density. The first high-latitude plate (*lower panel* (b)) appears very similar to the corresponding selection from the low-latitude field, with bright, blue stars overemphasized (also note that many of the very blue, $g - r < 0.2$ objects are gone after the first plate). By the second and third plates in this field the selected stars start to cover more of the parameter space; once a large fraction of the bright, blue stars have been assigned, many more faint, red M-type stars get selected.

M-type stars). Thus if one pre-selected three plates in a high-latitude field, the demographics of the stars on each observed plate would be quite different from each other. This makes reconstruction of the underlying populations rather difficult, because a different fraction of stars from each region of parameter space will have been observed depending on the stellar populations and stellar density in each field. Of course, variations in the number of times a given piece of sky is covered will dramatically alter the distribution of objects in the final catalog.

Finally, we note that a routine that weights stars for selection based on the local density in parameter space will produce catalogs with different target demographics for different regions of sky. This is inevitable, because as we just showed, the stellar density on the sky actually affects the distribution of selected stars in parameter space, such that there is no way to get identical samples from regions of sky with different stellar densities. We also note that for a survey such as LAMOST, with a circular field of view, it is not possible to cover the whole sky with each part sampled only once. This will inevitably make the sampling of certain regions of sky higher than others.

Thus, to determine the underlying stellar populations based on the observed spectroscopic sample, one would need to either simulate the entire selection process, or compare the number of spectra of each type observed in a given part of the sky to the number of that same type of star that was available in the photometric catalog. We note that holistic models of the Galaxy with tuneable analytic parameters are now available (e.g., the Galaxia code; Sharma et al. 2011) which could be sampled with the selection function of the survey and used to correct survey artifacts.

5 CONCLUSIONS

We have presented a general target selection algorithm that can be used in any instance where large numbers of stars are to be selected from a catalog that is much larger than the desired number of targets. The program performs selections in multi-dimensional parameter space defined by any number of observables (or combinations of observables). Various functions are available to emphasize certain types of targets, and the program can be readily modified to implement an overemphasis based on any smooth function of the observables. This target selection algorithm was developed for the LEGUE portion of the LAMOST survey, and has been implemented in the LAMOST pilot survey. We have shown that careful choice of the target selection parameters can produce the desired relative numbers of various target categories, while retaining a smooth distribution across parameter space.

References

Aihara, H., Allende Prieto, C., An, D., et al. 2011, ApJS, 193, 29
Allende Prieto, C., Beers, T. C., Wilhelm, R., et al. 2006, ApJ, 636, 804
Barden, S. C., Bland-Hawthorn, J., Churilov, V., et al. 2008, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 7014
Belokurov, V., Zucker, D. B., Evans, N. W., et al. 2006, ApJ, 642, L137
Carollo, D., Beers, T. C., Lee, Y. S., et al. 2007, Nature, 450, 1020
Chen, Y. Q., Zhao, G., Carrell, K., & Zhao, J. K. 2011, AJ, 142, 184
Chen, L., Hou, J., Yu, J., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 805
Cheng, J. Y., Rockosi, C. M., Morrison, H. L., et al. 2012, ApJ, 746, 149
Deng, L., Newberg, H. J., Liu, C., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 735
Dierickx, M., Klement, R., Rix, H.-W., & Liu, C. 2010, ApJ, 725, L186
Freeman, K. C. 2010, in Galaxies and their Masks, eds. D. L. Block, K. C. Freeman, & I. Puerari, 319
Freeman, K., & Bland-Hawthorn, J. 2008, in Astronomical Society of the Pacific Conference Series, 399, Panoramic Views of Galaxy Formation and Evolution, eds. T. Kodama, T. Yamada, & K. Aoki, 439
Grillmair, C. J., & Dionatos, O. 2006, ApJ, 643, L17
Klement, R., Rix, H.-W., Flynn, C., et al. 2009, ApJ, 698, 865

Lee, Y. S., Beers, T. C., An, D., et al. 2011, ApJ, 738, 187

Newberg, H. J., Yanny, B., Rockosi, C., et al. 2002, ApJ, 569, 245

Newberg, H. J., Yanny, B., Cole, N., et al. 2007, ApJ, 668, 221

Schlaufman, K. C., Rockosi, C. M., Allende Prieto, C., et al. 2009, ApJ, 703, 2177

Sharma, S., Bland-Hawthorn, J., Johnston, K. V., & Binney, J. 2011, ApJ, 730, 3

Smith, M. C., Evans, N. W., Belokurov, V., et al. 2009, MNRAS, 399, 1223

Smith, M. C., Whiteoak, S. H., & Evans, N. W. 2012, ApJ, 746, 181

Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645

Xue, X. X., Rix, H. W., Zhao, G., et al. 2008, ApJ, 684, 1143

Xue, X.-X., Rix, H.-W., Yanny, B., et al. 2011, ApJ, 738, 79

Yang, F., Carlin, J. L., Liu, C., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 781

Yanny, B., Newberg, H. J., Grebel, E. K., et al. 2003, ApJ, 588, 824

Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009a, AJ, 137, 4377

Yanny, B., Newberg, H. J., Johnson, J. A., et al. 2009b, ApJ, 700, 1282

Zhang, Y., Carlin, J., Liu, C., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 792

Zhao, G., Zhao, Y., Chu, Y., et al. 2012, RAA (Research in Astronomy and Astrophysics), 12, 723