Optimal Observing of Astronomical Time Series using Autonomous Agents

Eric S. Saunders

Submitted by Eric Sanjay Saunders to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Physics, September, 2007.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signed: ....
Eric S. Saunders

Date: ....

Abstract

This thesis demonstrates a practical solution to the problem of effective observation of time-varying phenomena using a network of robotic telescopes. The best way to place a limited number of observations to cover the dynamic range of frequencies required by an observer is addressed. An observation distribution geometrically spaced in time is found to minimise aliasing effects arising from sparse sampling, substantially improving signal detection quality. Further, such an optimal distribution may be reordered, as long as the distribution of spacings is preserved, with almost no loss of signal quality. This implies that optimal observing strategies can retain significant flexibility in the face of scheduling constraints, by providing scope for on-the-fly adaptation. An adaptive algorithm is presented that implements such an optimal sampling in an uncertain observing environment. This is incorporated into an autonomous software agent that responds dynamically to changing conditions on a robotic telescope network, without explicit user control. The agent was found to perform well under real-world observing conditions, demonstrating a substantial improvement in data quality for this mode of observing. As an example of an observing programme performed in the classical visitor mode of a non-robotic telescope, a survey of temporal stellar variability in the 13 Myr-old cluster h Persei is described. The inherent problems of such an observing programme are explored. Nevertheless, a significant fraction of the cluster members are found to be variable. Importantly, these variable members are found to lie almost entirely below the radiative-convective gap, suggesting a possible link between the change in energy transport in the stellar core and the topology of the surface magnetic field.

Contents

List of Figures

List of Tables

Declaration

This thesis contains work published or pending publication as papers. The results described in Chapter 2 will shortly be submitted as part of a paper to Monthly Notices of the Royal Astronomical Society. The main results described in Chapter 3 have been published in Astronomy and Astrophysics, Volume 455, pages 757-763, and in Astronomische Nachrichten, Vol.327, Issue 8, p.783. The description of the algorithm described in Chapter 4 has been submitted to Astronomische Nachrichten for publication in March 2008.

Acknowledgements

This thesis would not have been possible without the support and guidance of my supervisors Tim Naylor and Alasdair Allan. Tim has been responsible for most of what I know about astronomy, and his commitment to rigorous experimentation has been an ongoing inspiration to me. Alasdair's programming experience was invaluable to me in the final stages of this project, and his insight and help with the practical problems of the project were instrumental to its success.

As well as the helpful suggestions of my supervisors, the accuracy of the final manuscript owes much to excellent proofreading by Rob King and my father Bill Saunders, who identified innumerable mistakes, typos, and blatant leaps of logic.

For the h Per work, a big thank you goes to Nathan Mayne for his help and collaboration, and his insights into the vagaries of colour-magnitude diagrams. Thanks are also due to Stuart Littlefair for over-seeing the observing run where the data were collected, and for influential conversations about lightcurve analysis.

The agent-controlled experiment described in Chapter 4 would not have been possible without the help and support of the team at Liverpool John Moores University: Iain Steele, Steve Fraser, Chris Mottram, Neil Clay and Martin Burgdorf. Their good humour and prompt responses to the various crises involved in running a live experiment were instrumental to the results reported here.

Last but not least, a huge thank-you goes to my wonderful wife Rebecca. In the last few months, she effectively put her life on hold so that I could finish this work, and her constant and unwavering support are deeply appreciated.

Thank you all.

ESS
Exeter, Devon, U.K.
25th September 2007

“The process of going from an art to a science means that we learn how to automate something.”

Donald Knuth, *Computer Programming as an Art* (1974)

“Computer science is no more about computers than astronomy is about telescopes.”

Edsger Dijkstra

Chapter 1

Introduction

1.1 The rise of robotic telescopes

A simple definition of a robotic telescope is a telescope that can make observations without explicit human control. It is robotic in the sense that its low-level behaviour is automatic and computer-controlled. Such a telescope is typically run under the control of a *scheduler*, which provides high-level control by indicating astronomical targets for observation. The scheduler itself may be highly automated, as in the case of the Liverpool Telescope (Steele et al., 2004), using sophisticated optimisation algorithms to determine the best observations to make in real-time. Other robotic telescopes are scheduled in a simpler manner, and are subject to an external plan provided at the start of each night by a human astronomer. They are considered robotic only insofar as individual observations are carried out automatically. The privately operated commercial Tenagra II telescope[1] is a modern example of such a system.

The field of robotic telescopes has a long and venerable history. Around 1965, the Wisconsin Automatic Photoelectric Telescope, an 8-inch reflector coupled to a PDP-8 computer with 4KB of memory, was capable of making a number of operational safety decisions based on external sensor data. Once programed with the list of targets, the telescope was capable of running unattended for several days at a time, shutting down with the dawn or inclement weather (McNall et al., 1968; Code, 1992). At around the same time the 50-inch Kitt Peak Remote Controlled Telescope also came into operation, managed by a control centre 90km away in Tucson, Arizona. In this case scheduling was provided by a (graduate student) operator who would manually set the observation programme at the beginning of each night, and then periodically monitor the performance of the telescope as the run progressed (Maran, 1967).

The motivations that have driven the development of robotic observing systems are myriad. The pioneers of the field were motivated by a desire to improve observing efficiency and to alleviate the tedium associated with long term monitoring projects. Even in those early days, the benefits of real-time data processing were apparent. Colgate et al. (1975) used a microwave link to connect an automated 30-inch telescope at a remote mountain location with the significant computer resources located 30km away at the main university campus of New Mexico Tech. This

[1] `http://www.tenagraobservatories.com`

was driven by a need for real-time processing to identify supernovae quickly, in order to acquire follow-up spectra as early in the lightcurve as possible (Colgate et al., 1975).

1.2 Advantages of robotic telescopes

1.2.1 Speed

The theme of removing the human from the loop to allow faster observation response time has continued into the present day. Modern time domain astrophysics has successfully harnessed the speed advantages of robotic observing, perhaps most successfully in the study of γ-ray bursts (GRBs), where very early slews to target have led to significant advances in our understanding of these extreme events. For example Vestrand et al. (2005) were able to acquire unprecedented early lightcurve information for GRB041219a by automatically responding to an alert broadcast by the INTEGRAL satellite (Gotz et al., 2004), initiating the first exposure with their ground-based network RAPTOR within 8 seconds of receiving the message.

1.2.2 Automating the routine

Simple automation works well for telescopes with relatively static, long term observing programmes, such as dedicated survey instruments. One such is the Carlsberg Meridian Telescope (Helmer & Morrison, 1985), built in 1952 and initially operated in Denmark. In 1984 the telescope was moved to La Palma and fully automated. Driven by efficiency and cost considerations, the telescope was upgraded for remote operation in 1997. This setup obviated the need for an observer to be constantly present at the telescope, an unnecessary overhead for such a well-defined observing process (Evans, 2001).

1.2.3 Remote and extreme environments

If an observing site is very remote then there may be limitations on network bandwidth and site personnel which can make a robotic telescope desirable. The ROTSE-III telescope network comprises four telescopes strategically sited across the globe in Australia, Texas, Namibia and Turkey. Although situated close to existing observatories for maintenance purposes, the remote nature of the sites dictates an automated observation strategy with data reduction taking place at the telescopes themselves. Image files are stored locally until the disk is filled, when it is manually swapped out (Yost et al., 2006). This approach works particularly well when the science goals of the instrument are focussed enough to naturally give rise to relatively simple telescope schedules. In the case of both ROTSE and RAPTOR, alerts from the GCN gamma-ray burst network (Barthelmy et al., 2000) trigger override observations, enabling these telescopes to pursue their primary science objectives. For the vast majority of the time when a GRB is not occurring, however, these telescopes perform systematic sky surveys, building temporal maps of the sky for variability analysis, and looking for interesting or unusual transient phenomena (Woźniak et al., 2004).

The advantages conferred by making a remote observatory robotic are accentuated when that site is in addition located in an extreme environment. The Antarctic plateau is one such environment. During the winter months, the ambient temperature drops below -60°C, low enough to present significant difficulties for human instrument operation (Ashley et al., 2004). Under these conditions, a robotic telescope is likely to be cheaper, more reliable and more efficient than an equivalent non-robotic counterpart.

1.2.4 High volume observing

The sheer volume of observations that may be acquired by a relentless robotic observer such as RAPTOR marks a particular niche which has been exploited to perform novel science. The Thinking Telescopes project (White et al., 2006), an expansion of the original RAPTOR design goals, seeks to leverage the potential wealth of information buried in the dataset of night-to-night observations by applying artificial intelligence techniques to identify trends and anomalies in the data. In this case the project seeks to maximise scientific return on the instruments by deriving interesting secondary science from the data.

The SuperWASP project is an example of a wide-field survey instrument that takes advantage of high volume observations. It aims to detect transiting extra-solar planets by monitoring the varying fluxes of large numbers of bright stars ($V < 13$). Horne (2003) estimated that roughly 1000 hot transiting planets with masses of the order of Jupiter would be observable at a limiting magnitude of $V < 13$ (corresponding to a search depth of around 400 pc), and that wide angle planet searches could expect to discover 3–10 planets/month. Wide, shallow searches have the advantage that candidates are excellent targets for high-precision radial velocity followup, but the bright limiting magnitude dictates extremely wide fields of view, with only 2–3 objects expected per 10° by 10° field (Horne, 2003). The SuperWASP instrument comprises eight 200 mm camera lenses, each with a 7.8° by 7.8° field of view. Observations, including exposure, slew and readout times, take around a minute each and generate 8.4 MB of data per image (Smith et al., 2006). Thus it was a requirement that the entire observing process be largely automatic, including the use of a custom real-time data reduction pipeline to deal with the huge volume of incoming data.

Unfortunately, after the first observing season the project saw no detections with signal-to-noise > 10. This is believed to be due to the presence of correlated ‘pink’ noise in the dataset, which was not considered by Horne (2003). Simulations indicate that extending the time series of the same fields by another season will enable detections to be confirmed, albeit at a lesser rate than had been originally hoped (Smith et al., 2006).

1.3 The robotic observing paradigm

One interesting consequence of automating a telescope system is that the observing paradigm is altered. In a traditional observing run, an astronomer is allocated a block of time on a specific telescope. He travels to the remote telescope site and remains there for a period of some days, using the instrument to try to acquire the observations he requires. If the weather is poor, the seeing bad, or the observations interrupted by target of opportunity requests (typically overrides for fast transient phenomena such as supernovae or GRBs) then the astronomer is unable to use some fraction of his allotted time, and must make the best of what observations he was able to obtain. There is a finite risk that he will not accrue enough data of sufficient quality to adequately address his science problem. In this case he effectively leaves empty-handed. Regardless, the total length of the baseline of his observations may not be longer than the length of the allocated run.

In the robotic observing paradigm, these constraints are relaxed. Time on a network such as Robonet (Bode et al., 2004) is typically allocated per semester on the basis of scientific merit as a number of hours of on-sky time, not tied to any particular date. An observer submits an observation request or sequence of requests over the Internet and then simply waits. Whether or not an observation takes place is at the discretion of the controlling scheduler, whether this is implemented by a human or software. This has a number of consequences, described below.

1.3.1 Charging for time

Time is only debited if it is actually used. If an observation takes place, the time allotment of the user who requested that observation is debited accordingly. This means that as long as the observations are requested and scheduled before the expiry time specified by the telescope allocation committee (TAC), which is typically the end of a semester, then the *observations are implicitly guaranteed to take place*. The model does not explicitly manage the possibility of being systematically weathered out and thus leaving empty-handed. An alternative formulation of this principle is that *a user is only charged for the time they have used*. This change in the nature of the observing risk arises because the telescope scheduler has much greater flexibility with regard to placement of individual observations than in the traditional observing scenario. Because multiple users are scheduled in an interleaved fashion throughout the observing season, the risk of failures is spread across all users. This is particularly important because such failures are generally correlated in time, for example because of poor weather or instrument failure, which will render consecutive blocks of time unusable.

Note that while users gain by the guarantee of equitable charging, the service provider carries the cost of service downtime. In the traditional model, by the end of the semester all time will have been been accounted for, one way or another. Observations either took place, or they did not. In the robotic paradigm, a user with an allocated number of hours may make the reasonable assumption that their observations will happen at some point during the semester, and they will have a fair chance at the observations despite the presence of other, perhaps higher priority scientific programmes also allocated. Thus the provider must either ensure the load on the telescope closely matches the predicted availability of the service, or must explicitly disclaim any guarantee that the observations will actually happen, or must seek to manage the risk between allocated projects in some more explicit fashion. For example, *service mode observations* at the European Southern Observatory's (ESO) Paranal facility in Chile, while not technically robotic, share a number of similarities with the robotic mode. Around 60% of the observations at the Very Large Telescope (VLT) in service mode are performed by a support astronomer at the telescope without the presence of the principal investigator of the scientific proposal (Comerón et al., 2006). The likelihood that the observations will be performed is determined by the proposal's priority ranking, as defined by the VLT TAC. Observations with priority 'A' are guaranteed to take place if at all possible. Priority 'B' observations only take place if there is no higher priority observation that is applicable, but are considered likely to happen before the end of the semester (when the proposal will be terminated if not completed). Finally, category ‘C’ represents low priority observations. These generally have less stringent observational constraints and so may be used to fill gaps in the observing schedule. By pulling observations from these three pools as circumstances dictate, the schedule may be optimised towards the dual concerns of high scientific priority research and efficient use of the resource (ESO, 2004).

1.3.2 Non-consecutive observations

Since no observer need be physically present at the telescope, there is no requirement that observations occur in a single consecutive block of time [2]. This allows the somewhat peculiar scenario of an extended run length without a corresponding increase in the number of observations. If the run is concerned with acquiring signal information from a target, for example to determine the periodicity of a variable star, then an observer may take advantage of this feature of robotic observing to extend the baseline of the dataset, typically to obtain sensitivity to longer periods or to confirm a periodic modulation through repeated observations. However extending a run in this way reduces the fidelity of the sampling, and it is common for such observing runs to enter an undersampled regime, where the number of observations, if equally spaced across the duration of the run is less than the minimum number required for Nyquist sampling of the smallest period of interest. The practical question of how to deal with this situation is the subject of much of the present work, and is discussed in detail in Chapter 3.

1.3.3 Self-referential observing

An observer can at any time change his observing strategy based on the success or failure of observations that have been submitted to the robotic telescope(s) so far. Although the total amount of observing time available is normally static, the number of observations that may be *requested* is generally unlimited. Thus a valid strategy for a robotic telescope observer is to vary the request frequency dynamically in response to the success or failure of earlier observations. As we shall see in Chapter 4, this can be used to improve a user’s observing chances.

1.3.4 Multiple telescopes

Perhaps the most exciting feature of automated telescopes, whether fully robotic or operating in service mode, is that they exhibit emergent observational properties when a number of such telescopes are connected together in a network. For example, the diurnal sampling problems that plague single-site observations of periodic variables can be avoided if an observer can acquire data from a second, geographically distant site. Continuous monitoring projects such as the Whole Earth Telescope (Nather et al., 1990) and the BiSON network for solar observations (Chaplin et al., 1996) indicate the success of this approach. Another advantage of a robotic telescope network is

[2] Of course, there may be science requirements that dictate temporally close observations. For example standard star observations are typically made as close as possible to observations of the science target.

the ability to make contemporaneous observations, with both similar and different instruments (e.g. one might desire object spectra in addition to photometric measurements in several filters).

1.4 Disadvantages of robotic telescopes

1.4.1 Scheduling complexity

The main disadvantage of a robotic system is that automation requires work. The more sophisticated the degree of autonomy exhibited by the telescope, the greater the amount of work required to enable that functionality. A human astronomer performing a classical observing run on La Palma, for instance, has maximum control of the details of his observation. There is no scheduler to decide what he should or should not observe — his time has been block-allocated, and the only external scheduling consideration that could affect observations would be a rare target of opportunity override request. In a robotic system, the design and implementation of the scheduler is probably the single largest factor affecting performance. Dispatch scheduling, as used in state-of-the-art robotic systems such as the Liverpool Telescope (Fraser & Steele, 2004) and the STELLA robotic observatory (Granzer, 2004) is perhaps the most sophisticated scheduling approach attempted in astronomy thus far. Such systems typically seek to model observational parameters in terms of weighted merit functions, combining these functions in such a way as to enable an unambiguous best course of action for the telescope at any given time.

The simple notion of a 'best' course of action belies the deep complexity of this kind of problem. A robotic system is typically aiming to maximise a number of long-term goals, such as total observation throughput, fair allocation of observing resources amongst users, or completion of high-priority science programmes. But on an observation-by-observation basis, there are many short term goals that may be desirable: quality of data, for example, may depend on airmass and proximity of the target to the moon or other celestial object, as well as the current seeing. There are also operational constraints. Because slewing to a new target takes time, many schedulers implement some kind of slew-time merit, which penalises targets which are distant from the telescope's present position. This minimises the fraction of the time the telescope spends slewing (and not on sky) and helps to avoid pathological behaviour such as 'thrashing' of the telescope between widely separated targets. The general behaviour of most schedulers is governed by an *objective function*, which is a mathematical expression that encompasses the set of constraints the scheduler seeks to optimise.

More sophisticated scientific constraints may also exist when the observation is considered in the context of others in the same science programme. A monitoring campaign for a doppler imaging target requires dense, evenly distributed observations in rotational phase space (Vogt et al., 1987). On the other hand for a variability search where the rotation rate of the target is not known and there is a wide range of possible periods, a geometric spacing of observations provides the best way to evenly spread sensitivity to periods (Saunders et al., 2006b). These sorts of observing programmes are among the most challenging to implement successfully for dispatch schedulers.

More generally, a science programme which is nearing completion may be more of a priority, because significant resources have already been expended on the acquisition of this dataset, and it may be useless unless all observations are completed. Although superficially similar to the fallacious logic of the sunk cost fallacy (encapsulated in the idiom ‘throwing good money after bad’), this behaviour is in fact rational, if there are reasonable grounds for believing the programme can be completed, and if no alternative observing strategy would produce a higher return from the remaining time available. In terms of scientific payoff, the value of those final observations is much higher, given that the previous observations have been acquired. Handing in a thesis generates a large amount of value for the submitter, given the time invested in the research. On the other hand it may be that the data remain useful even though the dataset is (from a request standpoint) incomplete. Perhaps the observer asked for more time than was actually required, or part of the science question is still answerable. Adequately solving context-sensitive questions of this type remains a classic open problem in artificial intelligence (see for example Bratko, 1990, Ch. 14, and references therein).

Difficult questions also arise when observing programmes have relatively broad timing constraints. Consider the following scenario. A high-priority observation must be made. It is acceptable to perform the observation at any point in an eight-hour window, but the best location of the target in the sky will be near the end of the window. It is clear now, and the load on the telescope is relatively low. Should the telescope schedule the observation now, or wait and gamble on a potentially superior observation, risking possible bad weather or increased load?

The complexity of these issues has lead most scheduling implementations to make drastic simplifying assumptions in order to keep the real-world operational problem tractable. The Liverpool telescope scheduler, for example simply ignores long-term constraints. Whenever an observation can be made, the scheduler looks at the list of candidate requests, calculates the current observing merit of each potential pending observation, and selects the highest scoring observation. Fraser (2006) notes that although simple and robust, this type of simple dispatch model can only evaluate local optimisations, and there is no guarantee that the global solution arrived at by this method is in any way optimal. The STELLA system of Granzer (2004) provides some relatively sophisticated merit functions including phase coherence and an ‘after-pick’ merit for limiting cycles (these are analogues of the eSTAR S and N metrics described by Saunders et al. (2006a), and explained in more detail in Chapter 3). However the scheduler does not attempt to apply these merits itself — the parameters required for their correct formulation are left to be configured by the user. This makes the system very flexible, but at the price of requiring the user to supply optimised observation requests.

1.4.2 Interface limitations

One might argue that handing astronomers the responsibility for creating and managing favourable observing requests is reasonable. After all, astronomers are the experts with respect to their science, and have traditionally been entirely responsible for the choice and manner of their observations. However, in a dispatch scheduling context this is a subtle misrepresentation of the true situation, because full scheduling control has *not* been returned to the astronomer. Only high level decisions are off-loaded by dispatch schedulers; they are still responsible for determining all telescope-specific constraints, and only the telescope can know what the global request load looks like. In addition the priorities of the facility, while hopefully in sympathy with those of individual users, are not in general the same. Thus the interaction of the scheduler and the astronomer becomes the critical limiting point in the system. This is where the translation of what the astronomer needs into terms a scheduler can understand takes place. The level of sophistication of the system as a whole is largely driven by the richness of this interface. As a corollary, one could say that the degree of approximation that the interface enforces upon the precise communication of the scientific requirement to the observatory directly determines the sophistication of the observations that can be performed. An interface that only allows individual observations, or several equally spaced in a series, does not contain the basic grammar required to specify an optimal geometric series for undersampled variable star observations.

In the symbiotic ecology of robotic telescopes and their human observers a niche thus exists. While telescope systems evolve to better accommodate new types of observations and requirements, and astronomers learn what kinds of science are easy or well-suited to remote observing and bend their efforts in those directions, a third party able to communicate with the telescopes to provide a service that astronomers need is a useful enhancement to the system. Such a higher-level service enables the envelope of what is currently possible to be extended and tested without immediate large-scale infrastructure or sociological change. The eSTAR project fills such a niche. It aims to provide advanced client software to conduct some of the complex dialogues with telescopes that are required to do some of the most interesting science desired by the astronomer. Chapter 4 of this thesis concerns the implementation of one such system, an autonomous software agent capable of running an optimised variable star observation programme without human intervention.

1.5 Robotic telescope networks

1.5.1 Early networks

The idea of connecting a number of robotic telescopes together to create a global network is not new. As early as 1988, researchers had begun to consider the potential of a global network of robotic telescopes (Crawford, 1988). Crawford (1992) proposed the creation of a robotic network consortium, and offered a number of different potential use models. Drummond et al. (1995) prototyped a robotic telescope system called the Associate Principal Astronomer (APA), and claimed the first implementation of a new observing paradigm, in which astronomers submitted observation requests to a remote telescope by means of carefully crafted emails. The requests were written in a specially-designed instruction set called ATIS, which provided a rather cryptic ‘magic-number’ based syntax by which observations could be specified (Boyd et al., 1993). The language was rich enough to allow the specification of concepts such as observing groups and cadenced observations. The APA itself was conceived as a parsing and scheduling tool, that acted to filter the incoming observation requests and perform basic observation allocation. The system explicitly assumed a human support astronomer, who was responsible for handling the fine details of adjusting or improving the schedule, and of ensuring observation quality requirements were being met. However, lack of funding and the loss of key project members meant that the project stagnated, and never received wide-spread adoption. Actual implementation of a global network was to stagnate for the next ten years.

1.5.2 Homogeneous networks

It was not until the mid 1990s that the rate of appearance of robotic telescopes and semi-automatic, service mode telescopes began to accelerate (Hessman, 2006a). Decreasing computational and instrument costs, particularly CCD technology, meant that the construction of small automated telescopes to address specific science cases became increasingly feasible. This situation has lead to a not inconsiderable number of widely dispersed robotic telescope projects, each with access to small numbers of telescopes, most funded to execute specific science directives. At the top end of this group stand the research-grade robotic 2-metre observatories: the Liverpool Telescope on La Palma in the Canary Islands, and the two Faulkes telescopes, one located on the Pacific island of Hawaii, the other sited at Siding Springs in Australia. The smaller telescopes run the scientific gamut, with science goals that include GRB detection, occultation surveys, transit photometry and supernovae detection, among many others (Vestrand et al., 2002; Lehner et al., 2006; Christian et al., 2006; Quimby et al., 2006).

1.5.3 The benefits of networking

There are clear scientific advantages to joining a network as an individual telescope operator, or to connecting disparate project-specific networks together. Some of the advantages that can be leveraged include resource sharing, better coverage of time-domain observations, improved response times and follow-up capacity, and global efficiency gain across the network.

Naylor et al. (2006) described how the evolution of current robotic observing programmes would benefit from the introduction of heterogeneous observing resources, by allowing many different types of instruments to be brought to bear on the problem. Enabling telescopes with different instruments and capabilities to communicate amortizes the effective cost of specialised hardware, while increasing the range of science that can be performed with that hardware. This concept of adding value is the same basic mechanism that governs the principle of trade or barter of goods ("A voluntary, informed transaction always benefits both parties," Smith, 1776).

Certain science cases benefit from the unique advantages of multi-site observations. Observation programmes requiring very good phase coverage or those particularly sensitive to aliasing effects are greatly aided by the continuous coverage that a network can provide. Although a number of such networks exist for specific projects, a general purpose network would make this advantage accessible to any who could benefit from it, as a default property of the network. Thus periodicity measurements that might previously have been made from a single site by observers without dedicated instruments could be moved to the network, with substantial improvements in data quality at no additional cost.

A general network of telescopes is ideal for rapid followup of transient targets. A network with the capability to receive alerts from orbiting early-warning satellites such as Swift (Burrows et al., 2005) or specialist ground based instruments like the Large Synoptic Survey Telescope

(LSST) (Sweeney, 2006) can perform reliable, inexpensive followup observations with a large range of available instrumentation.

Networks of telescopes with even a limited ability to trade time are able to to offload or take on observations dynamically according to circumstances. This improves the global load-balancing behaviour of the network. This technique is very broadly applicable, arising in problems as diverse as bandwidth management of large computer networks, distribution of power across the electricity grid, and the logistics of supply chains for industrial goods (Foster & Kesselman, 1999; Østergaard, 2005; Erera et al., 2005). This potential for load-balancing is one of the ideas that has driven interest in the concept of formalising the notions of supply and demand in this telescope economy. Although several authors have proposed initial concepts for how such an economy might work (Etherton et al., 2004; Hessman, 2006a; Granzer, 2006; Steele et al., 2006), a number of significant issues have yet to be resolved. In particular, the metaphor of telescope time as a kind of currency, although appealing, may not be accurate. The basic notion is that an observer is allocated a given number of hours on a particular telescope network, and is then free to trade that time in any way he sees fit. There are several problems with this model. Unlike money, telescope time cannot be hoarded indefinitely. Neither does its value remain constant, or even approximately constant. An observer can convert generic 'time' into pending observations at any time, by making an observation request. Whether pending requests can or should be traded is an open question. Pending observation times that are far in the future are intrinsically less valuable than nearby times, which will be at a premium. Allocated time therefore becomes more valuable as the date approaches, but that value drops to zero if the date of the observation passes. Another issue is the problem of 'runs' on a particular 'currency': if everybody in the network wants one specific kind of time (perhaps an instrument or seeing constraint that is unique to a single telescope), then in principle they could all trade their 'local' currency for that time, and thus oversubscribe that resource at the expense of the rest of the network. This is a pathological case not only because it destroys the load-balancing properties of the network, but because as a result of the over-subscription the value of the traded time drops in real terms. The likelihood of an observation on this telescope succeeding falls (because it is too busy), and some number of 'customers' are likely to be dropped. These issues arise because the generalisation of 'telescope hours' is too broad, and there is an implicit assumption that if hours are available, they can be redeemed to produce real observations. Note especially that unlike money, these observations are not generic — they are specific. Eight hours of time on Robonet only has value if it can be redeemed for eight hours of observations of an object of interest, with the minimum observing conditions required for the execution of the science programme. Additionally, scientific value may be predicated on all the observations being performed, or observations placed at a specific cadence, and so on — the list is limited only by the creativity of the science cases that can be proposed. No satisfactory system has yet been advanced that can successfully accurately map resource availability to science problems, which is ultimately the only way to measure the success of the scheme. Fundamental work remains to be done in this area before any kind of market economics can be implemented successfully for robotic telescope networks.

1.5.4 Interoperability - building heterogeneous networks

A willingness to tackle the challenges and a shared desire for interoperability within the field lead to the formation of the Heterogeneous Telescope Network (HTN) consortium in July 2005. As well as providing a specialised forum for discussion and knowledge transfer between interested parties in the field, the HTN has three explicit aims (Allan et al., 2006):

1. Interoperability between robotic telescope networks.
2. Interoperability with the Virtual Observatory for event notification.
3. Establishment of an e-market for the exchange of telescope time.

The most common way of enabling interoperability between projects is through standardisation, and this is the primary focus of the HTN's work. Allan et al. (2006) defined a simple, implementation-neutral protocol that defines the interface that an HTN-compliant telescope is required to present to the network, and a standard set of transport protocols (Allan et al., 2008) are also specified. The standard was deliberately designed to be largely optional, in order to make the process of joining the network as easy as possible. A minimum HTN interface need only implement two phases of message dialogue: the exchange of the observation request and accompanying response, and the data return step. These mandate the use of a specific XML-based communication language, RTML (Pennypacker et al., 2002; Hessman, 2006b), which defines the set of standard terms required to describe telescope instruments and capabilities and to process observations.

1.5.5 Non-robotic networked telescopes

Although it is often generally assumed that a telescope network like the HTN should be based around robotic hardware to allow the system to behave in as automated a fashion as possible, this is not actually a *requirement*. In principle any dynamically scheduled telescope could accept observation requests or alerts from the network. This is particularly true for observatories that run primarily in service-mode, such as GEMINI and the VLT, or newer telescopes such as the 9 m Hobby-Eberly Telescope (Hill, 2000) and the 10 m South African Large Telescope (Stobie et al., 2000), which run entirely in service mode. The United Kingdom Infrared Telescope (UKIRT) (Humphries & Purkins, 1980) provides a good example of the integration of a large, 4-metre class research-grade telescope into a heterogeneous telescope network. A collaboration with the eSTAR project enabled the telescope to join the eSTAR network as an eSTAR observing node. In this case the non-robotic nature of the telescope is immaterial — the interface presented to the network is identical to other telescopes on the network. Specifically, observations can be scored, and then requested. At some later date they take place and the data are returned, or they are not made and expire. The details of how the scheduler and observation mechanism are implemented are of no concern to the network. In this example, an eSTAR agent placing a target of opportunity observation request with UKIRT enters the observation block into the queue database. It is then executed by the observer at the telescope, who is automatically compensated for the time by the dynamic scheduler (Economou et al., 2006).

1.6 Event notification

Event notification has been a field of growing interest within the time domain astronomy community. The γ-ray burst Coordinates Network (GCN) provides a fast, packet-based mechanism for distributing alerts from GCN sources to clients around the world interested in performing fast followup (Barthelmy et al., 2000, 1994). One of the main functions of the GCN servers is to broadcast event information triggered by dedicated space-based alert systems such as the Swift satellite.

Event-driven observing is not limited to GRBs, however. The Supernova Early Warning System (SNEWS) provides notifications of supernovae events detected by neutrino experiments such as Super-Kamiokande. The approach takes advantage of the fact that the neutrino flux from a collapsing stellar core emerges some hours before the corresponding photon signal (Antonioli et al., 2004). There are a number of other project-specific event mechanisms in use by specific groups (White et al., 2006).

Although they have enabled important new science, these systems nevertheless have some significant limitations. While the alert notifications emitted by the GCN are rigidly formatted and therefore easily parsed, GCN email circulars generated by clients, which often contain refined coordinates, are plain text, natural language messages not designed for computer consumption, and this has handicapped automated indexing and searching efforts (White et al., 2006). This is a particular concern for projects that wish to leverage the data mining potential of historical databases of such messages, for example to identify recurring or persistent trends associated with particular sources (Vestrand et al., 2004). These difficulties led workers within the Virtual Observatory community to seek a way to define a standardised message format, leading to the formation of the VOEvent working group in early 2005 (White et al., 2006).

The VOEvent notification format (Seaman et al., 2005) was born from a single clear vision: to take the best feature of the existing mechanisms, namely the concept of simple, timely, useful messages, and create a generalised formalism that could be easily parsed by machine. By using XML (Bray et al., 1998) to specify the syntactic relationships between concepts, it was possible to satisfy the desire for human-readability while enforcing the requirement that the metadata be exact. The format aims to provide a concise description of the critical parameters required to describe an astronomical observation. Conceptually, these can be thought of as answers to the six basic journalistic questions: who, what, when, where, why and how (White et al., 2006; Kipling, 1902).

For the significant fraction of the robotic telescope community who rely on such alerts, a precise, transparent, searchable format for event messages is a great advantage. However, the relevance to an HTN in general is more subtle. Hessman (2006a) pointed out the architectural similarities between a system in which events are produced and consumed, and one in which observations requests are produced and resulting observations consumed. Indeed, the two systems of events and observations closely co-exist. Producers and consumers of both kinds form intertwined logical networks mapped onto the physical topology of the telescopes themselves. Events in themselves make no demands of any actor on the network. However, because entities interested in making observations often do so on the basis of the events they receive, they effectively drive some subset of the observing behaviours that take place on that network. At the same time, the ability to generate alerts provides the potential for active collaborations and opportunistic behaviour which could enhance the efficiency of the network as a whole. For example, a GCN alert triggers observations within the RAPTOR sub-network. RAPTOR then broadcasts the details of its observations to collaborators as VOEvent messages. An eSTAR agent with time on a different network determines that followup observations with a particular instrument would be helpful. Having successfully requested and obtained the data, an event message is broadcast, which may lead to a change in the observational priorities of the original RAPTOR network in turn.

1.7 The future: scaling the network

The robotic telescope projects discussed each control a small number of telescopes (typically less than five). The emergence of the Las Cumbres Observatory Global Telescope (LCOGT) project marks a significant step change for the field. LCOGT is a privately funded initiative to build a global network for performing science as well as education outreach activities. It will comprise a core based on two of the former Robonet 2 m telescopes (the FTN and FTS), and a widely distributed secondary group of smaller robotic telescopes of varying sizes. The ambitious aim of the project is to perform novel science that utilises the potential of a true distributed network (Rees et al., 2006). Because the telescope is operated and managed by a single administrative entity, the sociological problems faced by groups such as the HTN in relation to the exchange of telescope time do not apply internally to this network. Instead, the major outstanding challenge is how to coordinate and schedule the network in such a way that the complex global behaviour of the system is well-optimised. This question is multi-layered and complex, since the ultimate operation of the network is fundamentally defined by the nature of the instrumentation and control software, the network architecture, the data reduction, processing, storage and transfer, as well as the higher-level problem of making decisions about what should be observed, where and when. Nevertheless, for the foreseeable future, the questions of why and how observations should be performed remain the province of astronomers.

1.8 Thesis summary

This chapter has described the use of robotic telescopes from their inception to the present day. Some of the unique advantages and possibilities enabled by this technology, as well as the specific challenges inherent in these systems have been examined. In particular the problem of telescope scheduling has been discussed. Finally, a brief appraisal was made of the state of the art in the field, namely the linking of multiple telescopes to create observational networks for particular science goals, and more recently, the first steps towards the creation of truly global heterogeneous telescope networks.

This sets the scene for what follows. This thesis is concerned with the exploitation of the unique environment of a modern robotic telescope network to facilitate time series observations. Chapter 2 presents a variability analysis of the young stellar cluster h Persei obtained in the traditional, non-robotic manner. The dataset was time-consuming and costly to obtain, and suffers from classic problems such as diurnal aliasing, the inevitable consequence of constraining all observations to take place from a single geographic location. It serves as a typical example of the type of observing programme that this thesis addresses with much greater efficiency in the robotic paradigm, and without recourse to human interaction. The data are scientifically interesting in their own right, however, and so a detour is made to consider the implications that the distribution of temporal stellar variability as a function of colour and magnitude has for models of the stellar interior, in current theories of pre-main-sequence stellar evolution.

Chapter 3 asks a deceptively simple question: What is the best way to make time series observations, if the total observing time is constrained, and sensitivity to a range of periods is a scientific requirement? In particular, what is the best strategy for an observer using a robotic telescope network, no longer compelled to complete all observations within a single contiguous observing block? The exploration of this question begins with an analysis of what makes one time series better than another. A set of tools for inspecting the quality of time series datasets are developed, and used to demonstrate empirically the existence of datasets with superior period sensitivity when compared to equivalent datasets with the same number of observations, the same baseline, and the same total observing time. It is shown that the *temporal position* of observations in the dataset is critical to the ability of that dataset to recover signals of different frequencies effectively. An explanation in terms of minimisation of structure in the sampling window function is presented. A simple prescription is then developed that allows the generation of a set of optimal observation timestamps, as a function of the period range of interest and the total number of observations to be made.

Chapter 4 solves the practical problem of applying the theoretical insights described in Chapter 3 in an automated, reliable way in a true production environment, that of a professional robotic telescope network. The notion of an autonomous agent, a computational entity that manages an observing run independently of an astronomer, is explained and described in the context of the eSTAR multi-agent system. Some of the pitfalls that await the unwary observer in the hostile, self-oriented environment of a modern robotic telescope network are described. Based on these concerns, an algorithm is developed that allows an agent to flexibly respond to changing observational conditions, continuously optimising the choice of observation placement in order to maximise the quality of the final obtained time series. The agent was used to autonomously manage an observing run targeting the variable star BI Vir, and its performance is evaluated in the final part of the chapter.

Finally, Chapter 5 pulls together the disparate threads, summarises what has been achieved, and presents the conclusions that may be drawn from this work.

Chapter 2

Variability in the Young Cluster h Persei

2.1 Introduction

In this chapter, a survey of temporal stellar variability in the h Persei cluster is described. The double cluster h and χ Persei is a young, bright open cluster that has been the subject of detailed observations for seventy years (e.g., Oosterhoff, 1937; Crawford et al., 1970; Tapia et al., 1984; Waelkens et al., 1990). More recently, attention has been focused on mass segregation (Slesnick et al., 2002; Bragg & Kenyon, 2005). The cluster is rich in pre-main-sequence (PMS) objects, and is therefore an important target for observers interested in star formation and accretion processes. With respect to this thesis, the main purpose of this chapter is to describe in detail the process by which an astronomer in the classical, single-telescope visitor mode paradigm proceeds to acquire observations, reduce these data to useful lightcurves, and then extract useful scientific value. This is the principal use case for which the work of Chapter 3 was originally inspired, and the underlying driving motivation for the software agent described in Chapter 4. Although the work in these subsequent chapters is generally applicable to a much wider range of time-domain studies than this single science problem, it is nevertheless instructive to pursue this example, to thoroughly understand the scientific requirements, and also to identify improvements that can be made through the use of a robotic network and automated observing tools.

A very brief overview of star formation is presented in Section 2.2. This is followed by the details of the observing run (Sec. 2.3) and the process of data reduction (Sec. 2.4). Section 2.5 describes the way that variable stars were identified and selected from the overall mass of observed cluster stars. Section 2.6 presents the results, comparing the variable population with the cluster as a whole in terms of distribution in a colour-magnitude diagram (CMD). A discussion of the results and their relevance to the current state-of-the-art in star formation is presented in Section 2.7. Finally, the chapter is summarised and conclusions presented in Section 2.8. This chapter is based on Saunders et al. (2008, in prep.).

2.2 A brief overview of star formation

Stars are formed within massive clouds of molecular hydrogen, which are relatively dense in comparison to the surrounding interstellar medium. The clouds themselves often appear opaque, their dusty composition absorbing the light from background stars. As these ‘dark clouds’ disperse, they reveal many visible young stars, the youngest of which are the T Tauri stars, the subject of this chapter.

The process of forming such a star is remarkably complex. The standard view is that the basic physical process is the accumulation of interstellar gas and dust under the attractive force of gravity. This idea was first articulated in 1692 by Newton (Newton, 1756). Kant (1755), based on the work of Swedenborg (1734), developed the so-called *nebular hypothesis*, where he proposed that the action of gravity upon a slowly rotating cloud leads to the formation of a disc, eventually giving rise to a star and planets. Laplace (1796) independently developed a similar idea. A theoretical basis for initiating such a collapse, called *gravitational instability*, was provided by Jeans (1902), who showed that small density perturbations in an initially uniform medium lead to runaway gravitational collapse.

There are two main models for the collapse to form the initial cloud core, which represent extremes of fast and slow collapse respectively. In the rapid formation model, initial instability allows gravity to overcome thermal pressure, leading to runaway collapse (Hayashi, 1966). Alternatively, the slow collapse model of Shu (1977) considers a core supported by a magnetic field, which accretes matter gradually through ambipolar diffusion. It is likely that the true situation lies somewhere between these two extremes (Larson, 2003).

In either case, the density continues to rise, but the temperature remains almost constant at around 10 K due to thermal coupling of the gas to the dust, a situation which continues whilst the core remains optically thin. As the core collapses, a pressure gradient is created, because the central region is growing in density while the other regions are not. This outward pressure gradient ensures that most of the initial mass remains in the extended envelope, and the resulting protostar formed at the centre of the collapse has a very small initial mass. Rotation, while helping to ensure that most of the remaining mass joins an accretion disc around the new protostar, is not sufficient to prevent ongoing growth of the central density singularity.

Eventually the core reaches a critical density of 10^{-13} g cm^{-3}, and becomes opaque to thermal radiation. This leads to a rapid rise in temperature. The core can no longer be considered isothermal, and enters an adiabatic phase, halting collapse as the pressure exceeds gravity. This is the formation of the first hydrostatic core, with a mass of about $0.01 M_{\odot}$ and a radius of several A.U. The stability is only temporary, however and the protostar continues to gain mass through accretion.

Dissociation of hydrogen molecules can take place once the temperature exceeds 2000K, and the collapse resumes, because energy goes towards the dissociation process rather than heating. This continues until the hydrogen has ionised, bringing the collapse to a halt with the formation of the second hydrostatic core. The protostar continues to accrete matter, depleting the surrounding envelope which becomes optically thin. This allows the protostar to radiate energy freely, enabling a constant radius of about $4R_{\odot}$ to be maintained. At this point deuterium fusion begins. The protostar is a fully-fledged pre-main-sequence star, with only a relatively small amount of matter remaining to accrete.

While obscured by dust, the new ‘young stellar object’ (YSO) is only observable in the infrared or sub-mm. Once most of the accretion is completed, the observed spectra of the object is a combination of infrared emission from the surrounding disc and visible light from the star itself. Finally, once the disc is gone, the light moves entirely into the visible spectrum. These observational properties have lead to the designation of a set of classes of YSO, corresponding to their observational properties. Class 0 objects are in the stage of rapid early accretion (lasting about 10^4 years), and are only observable in the sub-mm. Class I objects are in the main accretion phase, which takes about 10^5 years, and are visible in the far infrared. Class II objects are classical T Tauri (CTT) stars whose spectra are a composite of the star itself and an attendant dusty disc, visible in the near infrared. This stage typically lasts for around 10^6 years. Finally, class III objects are the fully visible weak-line T Tauri (WTT) stars, whose discs are optically thin.

The initial radius for all low to mid-mass YSOs is approximately the same. On a Hertzsprung-Russell (H-R) diagram of temperature vs luminosity, this leads to a feature known as the *birthline*, representing the range of points at which a newly observable star can enter the H-R diagram. From its arrival on the birthline, a pre-main-sequence star will undergo a period of contraction of around 10^7 years, before the onset of hydrogen burning. This marks the end of the pre-main-sequence stage of the star's life, and the beginning of its progression along the main sequence.

2.2.1 T Tauri stars

T Tauri stars are low mass, pre-main-sequence stars. In comparison to main-sequence objects of equivalent mass, T Tauri stars appear over-luminous due to their larger radii. As previously described, these stars slowly contract as they move towards the main sequence on the H-R diagram, and it is this gravitational contraction, rather than hydrogen fusion, that provides their radiative energy source. Surveys indicate that around half of all pre-main-sequence stars possess dusty circumstellar discs (Strom et al., 1989; Beckwith et al., 1990). A dusty disc absorbs light from its parent star, re-emitting it in the infrared, and this can explain the spectral signatures of many CTT stars (Kenyon & Hartmann, 1987). Some CTT stars possess infrared components larger than can be explained by disc re-emission. This 'infrared excess' can be explained by accretion infall (Lynden-Bell & Pringle, 1974).

WTT stars exhibit coronal emission similar to that of the Sun, but at much higher levels, indicating strong surface magnetic fields (Hartmann, 2001). These magnetic fields hold the accretion disc away from the stellar surface, but allow accreting material to be funnelled down to the star along magnetic field lines, producing hot spots at the shock interface between the accreting material and the stellar surface (Köenigl, 1991; Hartmann, 2001). The rotation of the star causes these surface structures to appear and disappear from view, leading to a visible modulation in the apparent brightness of the star with period equal to the rotation period (Bertout et al., 1988). This is believed to be the principal cause of observed variability in T Tauri stars.

2.3 Observations

The observations discussed in this chapter were performed by Stuart Littlefair and the author, using a Sloan *i* filter (Fukugita et al., 1996) with the 2.5 m Isaac Newton Telescope (INT) on La Palma, using the Wide Field Camera with the four CCD EEV array. The run spanned 16 consecutive nights, from the 22nd September to the 7th of October, 2004. There were two photometric nights, the 27th September and the 5th October. h Per was observed on 12 of these nights, for around 2.5 hours per night. Exposures of 5, 30 and 300 seconds were obtained. There were 110 good frames taken for the 5 s and 30 s datasets, and 213 frames for the 300 s dataset.

2.4 Data reduction

The data were reduced, and photometry extracted using the optimal extraction algorithm of Naylor (1998) and Naylor et al. (2002), with the modifications introduced by Littlefair et al. (2005). The steps required to proceed from the set of observed frames to the set of lightcurves for each observed star are described here. Broadly, the data reduction consists of the following processes.

1. Bias subtraction, flat-fielding and defringing.
2. Offset determination.
3. Object detection and quality determination.
4. Optimal photometry.
5. Profile correction.
6. Relative transparency correction.
7. Astrometric solution.
8. Normalising instrumental magnitudes to produce the final catalogue.

These steps are described in detail below.

2.4.1 Initial stages

The images were bias subtracted using a single master bias frame, corrected for flat-field effects using a master twilight sky flat-field constructed from images obtained for each night of the run, and defringed by median stacking many frames to produce a master fringe frame. At this point the images are ready for offset analysis. The aim is to determine the pixel offset and rotation (if any) between each frame and a chosen reference frame.

Stars to be used as reference points are identified by successive passes of a peak finding algorithm that identifies pixels with significant electron counts. Beginning with an initial pass that identifies the highest unsaturated values in the image, and dropping the significance level with subsequent passes enables a set of probable stars to be determined. By fitting an estimated point spread function (PSF) based on the star with the highest non-saturated pixel value, rough positions for the stars are determined. Magnitudes are determined from the fitting, by considering the full width at half maximum (FWHM) of the PSF. The mapping between the reference frame and the remaining frames is then determined by comparing the pattern of identified stars in each frame. The result is a set of pixel offsets and rotation angles for each observed frame.

Before moving on to optimal photometry, more accurate star positions need to be determined, and a list of stars suitable for use as PSF indicators obtained. This is achieved in the following way.

The level of background counts due to the sky (the 'sky level') is determined by measuring the modal counts in a series of sky boxes, chosen to be large enough that the presence of a star within the box would not have a large impact on the calculated sky level. This sky level is then deducted from each pixel, and small variations in the remaining background smoothed by passing a top-hat filter over the image. In a similar way to the star detection algorithm used to determine the offsets between images, a series of passes of decreasing significance is made over the image. At each stage, significant peaks are tracked and a two-dimensional connectivity condition applied to identify adjacent pixels belonging to the same peak. In this way individual pixel values are located as belonging to particular stars.

A number of data quality checks are made during this process, including the identification of duplicate detections of the same star, and the flagging of stars that contain bad or saturated pixels, pixels with counts in the non-linear response range of the detector, or pixels with negative counts. Stars possessing a PSF that is not point-like are flagged for non-stellarity. This could be due to the presence of another star nearby in pixel space, or because the target is an extended object such as a galaxy or man-made object. Additional flags may be applied to stars that fall close to the edge of the CCD or on a known bad sector of the detector.

The result of the object detection stage is a list of target positions, flagged by quality. Stars with no flagged problems are candidates for PSF template selection, to be used in the accurate photometry described in the next section.

2.4.2 Optimal photometry and profile correction

Optimal extraction is a technique for maximising the signal-to-noise ratio of a given star. A detailed description of the method is outside the scope of this chapter (see Naylor (1998) for the full details), but the key idea is to treat each individual pixel as an independent estimator of the flux in the stellar profile. This is achieved by normalising the signal in each pixel by the fractional flux predicted in that pixel by the PSF. Independent flux estimates from each pixel are then optimally combined by weighting them as a function of their uncertainties. The technique therefore requires a PSF to be estimated. The closer this estimate is to the true PSF for the star, the higher the signal-to-noise that can be recovered.

The brightest stars from the PSF candidate list are resampled to the pixel grid of the star of interest by fitting a two-dimensional Gaussian to the observed PSF. The PSFs are ordered by the size of the corresponding FWHM, and the median of this list is selected as the model PSF. This provides a simple, automated way to select a reasonable estimator PSF for the optimal weighting. Fluxes are then extracted for each star, by taking the geometric mean of the PSF in each axis and integrating under the curve.

The extraction mask derived in the optimal photometry stage is likely to be close to the true

PSF of the target star, but in general is not exact, since the true stellar profile is not an analytic function (see for example Diego, 1985). The profile correction is the logical analogue of an aperture correction in standard aperture photometry, and empirically corrects for the differences between the extraction mask and the true PSF. A number of unsaturated stars in each CCD provided the profile correction for that CCD, which is a polynomial function of position. It is measured by comparing the flux measured for the objects using the optimal extraction with that obtained by applying a somewhat larger aperture (although in this case the crowded field limited the effective aperture size). Note that, because in general the correction is a function of position on the detector, this correction is still important, even though the photometry in this case is relative, and not tied to any particular system (see e.g. the discussion in Littlefair et al., 2005).

2.4.3 Final stages: transparency correction and astrometry

A relative transparency correction was applied to all the resulting photometric measurements, to normalise any differences between frames arising from variations in airmass or transparency. This was achieved by selecting a subset of relatively bright, unvarying stars and using their average magnitude to define a reference brightness for each frame. This was done using an iterative procedure to identify the least varying stars based on computation of the reduced χ^2, denoted χ^2_ν, in the same way as described in Littlefair et al. (2005).

To perform astrometry for each object, for each exposure time (5 s, 30 s or 300 s) a single frame was chosen, and the combination of the three then searched to create a master catalogue. The objects were then matched to a 2MASS catalogue of the field to produce an astrometric solution, providing a transformation from pixel coordinates to J2000 equatorial coordinates, with a mean RMS discrepancy in positions of 0.1 arcseconds.

2.4.4 Normalising the combined final catalogue

An arbitrary zero point was applied to the instrumental magnitude so that the observations would be in a natural magnitude system approximately comparable with Sloan i. Magnitudes in this system are denoted by the symbol i_I. Stars with mean i_I magnitudes brighter than 17.5 but fainter than 16.5 were recorded by both the 300 and 30 s datasets. Since the photometry for each dataset is only relative within the frame, the 5 s and 30 s datasets were renormalised relative to the 300 s dataset so that the three datasets could be combined into a final catalogue. This placed all three datasets on the same scale in i_I. It was then possible to pick limiting mean magnitudes for each dataset. The 5 s dataset was used for all stars brighter than an i_I of 15.2. For the overlap region between the 30 s and 300 s datasets, a delineating i_I magnitude of 17.5 was chosen, so that stars with mean magnitudes brighter than 17.5 were taken from the 30 s dataset, while stars fainter than an i_I of 17.5 were taken from the 300 s dataset.

The ultimate result of this data reduction process was a set of lightcurves for all the stars in the field, amalgamating individual flux measurements from all three exposure sets. The work of examining these lightcurves for potential variability could now begin, and is described in the following section.

2.5 Variable star selection

The identification of variable stars involved a set of distinct selection criteria. First, the dataset was stripped of stars flagged as problematic during the data reduction process (see the discussion in Section 2.4.1 for details of the flagging process). Due to high vignetting a number of spurious variables at the corner of CCD 3 were also removed, using a simple coordinate cut. Poor quality objects were then removed by applying a selection cut at a mean signal-to-noise (S-N) of 10 to each lightcurve.

Having dealt with the most obviously problematic objects, the next stage was to evaluate the stars at the lightcurve level, to ensure a minimum number of flux measurements had been made for each potential variable star. A requirement was set of a minimum of 60 good datapoints from the 5 and 30 second datasets (from a possible 110), and 120 good datapoints from the 300 second dataset (from a possible 213), respectively. These limits were chosen by considering the theoretical probability of χ^2_ν for different numbers of observations. χ^2_ν was calculated by fitting a constant value to each lightcurve. For more than 120 observations of the 300 second dataset, the probability of exceeding a given χ^2_ν was found to fall within a narrow band, indicating a suitable detection threshold. A similar detection threshold was found for the 5 s and 30 s datasets with a minimum of 60 observations. Although it would have been possible to choose the χ^2 threshold to be constant, this would have presented a problem, because the quality of the data is lower at higher points on the sequence. Thus a constant χ^2 threshold would produce a discontinuity at the join between any two of the datasets. The combined CMD indicates that this is not a problem with this choice of χ^2 (see Fig. 2.1).

The optimal photometry measurements provided an uncertainty for each datapoint in each lightcurve. Using these, the ratio of the mean flux to the uncertainty on every datapoint was calculated for each lightcurve. Since this was an attempt to sample an expected constant value, an RMS was defined in terms of the weighted mean of the dataset. The datapoint with the highest value of the ratio of the flux to the uncertainty was discarded, and the remaining datapoints used to find the weighted mean. This improved the ability of the subsequent χ^2_ν test to detect true variability by eliminating spurious cosmic rays or other oddities from the data.

Following Littlefair et al. (2005), a threshold of $\chi^2_\nu > 10$ was fixed as a cutoff indicating significant variability in the 300 s dataset. The correct χ^2_ν limit for the 30 s dataset was then determined by considering stars falling within the overlap region between the two datasets. Varying the value of χ^2_ν applied to the 30 s dataset, it was found that a minimum threshold of $\chi^2_\nu > 3$ recovered a number of variable stars in this region that was comparable to the number of stars detected in the 300 s dataset. Similarly, by considering variables identified in both the 5 s and 30 s dataset, it was determined that a threshold of $\chi^2_\nu > 3$ applied to the 5 s dataset recovered all variables to $i_I \approx 15.2$. Magnitudes and colours were obtained by cross-matching against the h Per catalogue of Mayne et al. (2007), to obtain a CMD for the variable stars in the optical bands V and V-I (Fig.2.1)

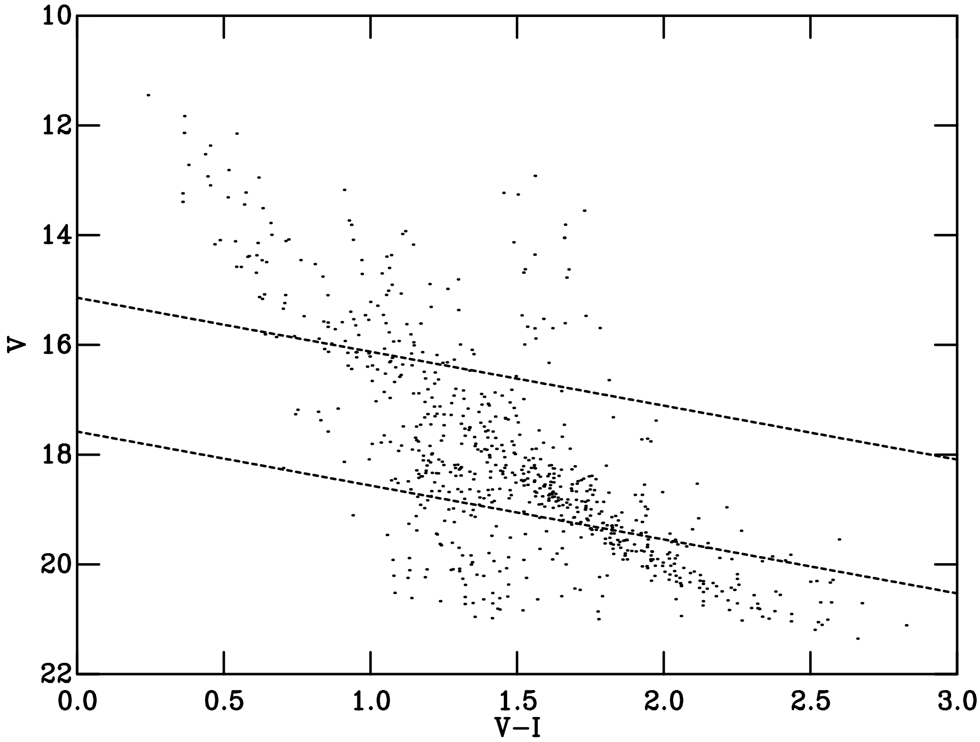

Figure 2.1: CMD of the variable population, after combining the 5 s, 30 s and 300 s datasets. The dashed lines indicate the limiting magnitudes used to combine the datasets. The unbroken continuation across the dataset boundaries indicates the selection criteria are not a function of sensitivity.

2.6 Results

Figure 2.1 shows a clear pre-main-sequence population running diagonally from approximately 14th to 20th magnitude in V. However, background contamination is also evident, and is clearly visible in Fig. 2.2 (upper panel) where it is concentrated into a large, wedge-shaped area that cuts laterally across the sequence. One common problem in this type of study is the presence of contamination objects in the sample, which can obscure the population of interest. These are additional stars in the observed region that are either field stars, unassociated with the target cluster, or stars of other types that may fall in similar regions of the CMD to the T Tauri stars. Below the sequence at a V-I of 1.0–1.5 contamination is likely due to dwarfs, the tip of which carries on above the sequence to around 13th magnitude. Additional contamination at a V-I $\approx$ 1.6 is probably due to background giants. In order to improve the membership fraction of variables for analysis a 10 arcmin radius circular cut was applied, centred on the h Per centre-point at α = 02 18 58.76, δ = +57 08 16.5 (J2000). Although the size of this cut is rather arbitrary, it was found that the results were relatively insensitive to the choice of radius. Figure 2.2 (lower panel) shows the variables remaining after this process. In order to trace the variables as a fraction of likely h Per members at different points along the cluster sequence, the same circular cut was applied to the h Per catalogue of Mayne et al. (2007) (Fig. 2.2, upper panel). Identical areas were deliberately selected to ensure that no bias would be introduced by possible mass segregation.

Looking at Figure 2.2 (lower panel), it can be clearly seen that the density of stars falls dramatically as we progress up the sequence. This is quite different to the distribution of cluster members (Fig. 2.2, top panel), which shows a strong sequence up to around 15th magnitude.

2.6.1 Histogram analysis

Although it is clear from visual inspection of the sequence that the density of variable stars in the CMD falls dramatically at brighter magnitudes, additional analysis is required to investigate this behaviour in a robust way. The idea is to quantify the prevalence of variables at different points in the sequence. In order to be meaningful, this has to be done with respect to the orientation of the sequence itself. To examine the distribution of variables along the cluster sequence, histograms are constructed according to the following scheme. A straight line was fitted to the central portion of the sequence, running from 1.0 to 2.0 in V-I. Two cuts perpendicular to this sequence line were then taken, a bright cut crossing the sequence at 1.0 in V-I, and a faint cut crossing at 1.66, with widths of 0.4 in colour-magnitude space. The stars selected by these cuts were then pulled back onto the cut centre-lines and binned in V-I. Fig. 2.3 illustrates the cut placement for the variable stars. In order to remove contaminating field variables, the number of variables falling outside of the 10' radial cut in each colour bin were scaled to match the area of the cut, and then subtracted from the number of variables within the cut. Any remaining overdensity is then taken to represent the cluster variable population. These cuts produce the histograms of Figure 2.4.

The histograms show in detail what can be seen qualitatively by eye in the CMDs of Fig. 2.2. While the general cluster membership continues to relatively bright magnitudes, the variability dies out at magnitudes brighter than about 17. In the left panel of Fig. 2.4, which shows the faint

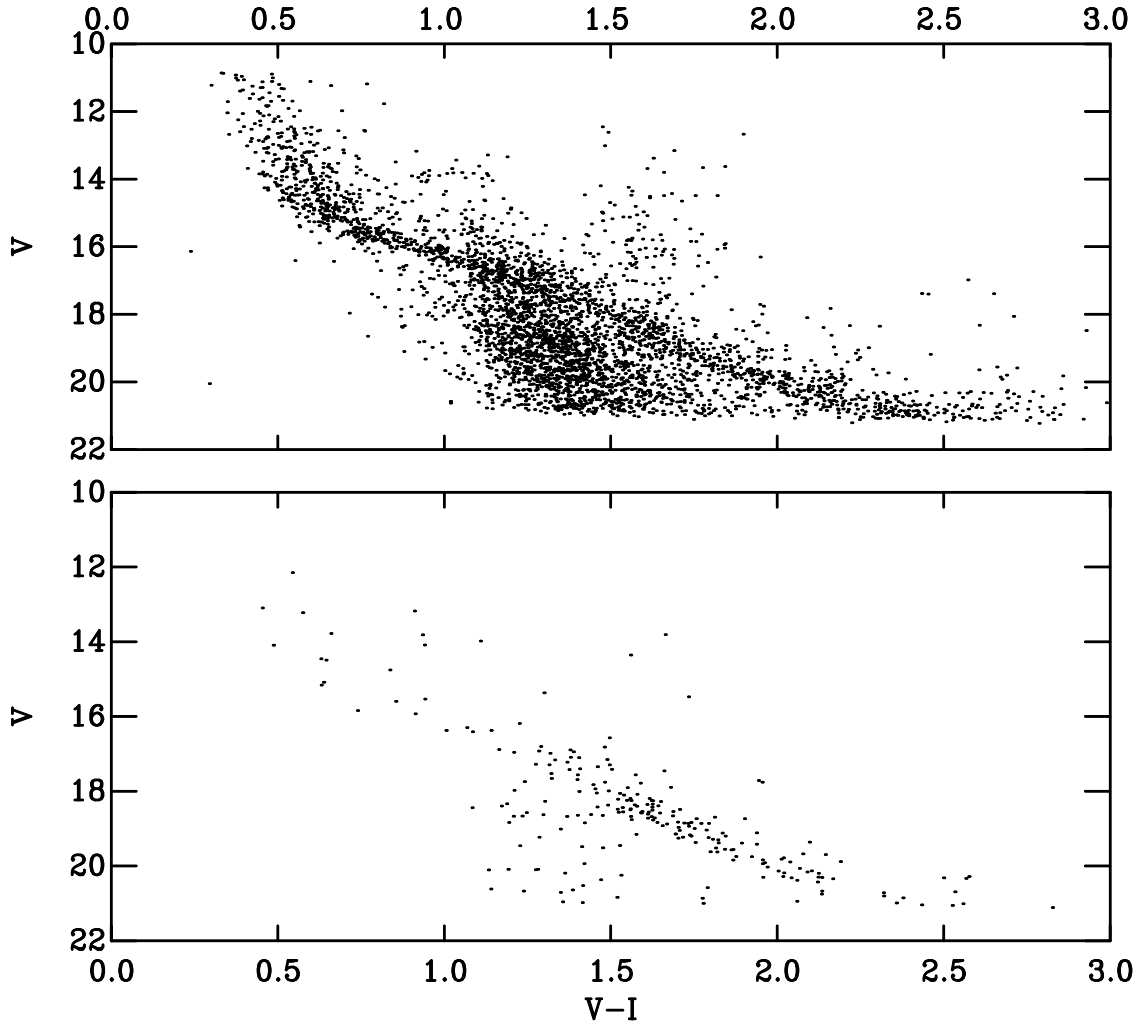

Figure 2.2: CMD of probable h Per members taken from Mayne et al. (2007) (upper panel) and variables (lower panel) after a 10' radial cut centred on h Per. The number of variables drops at V-I < 1.5.

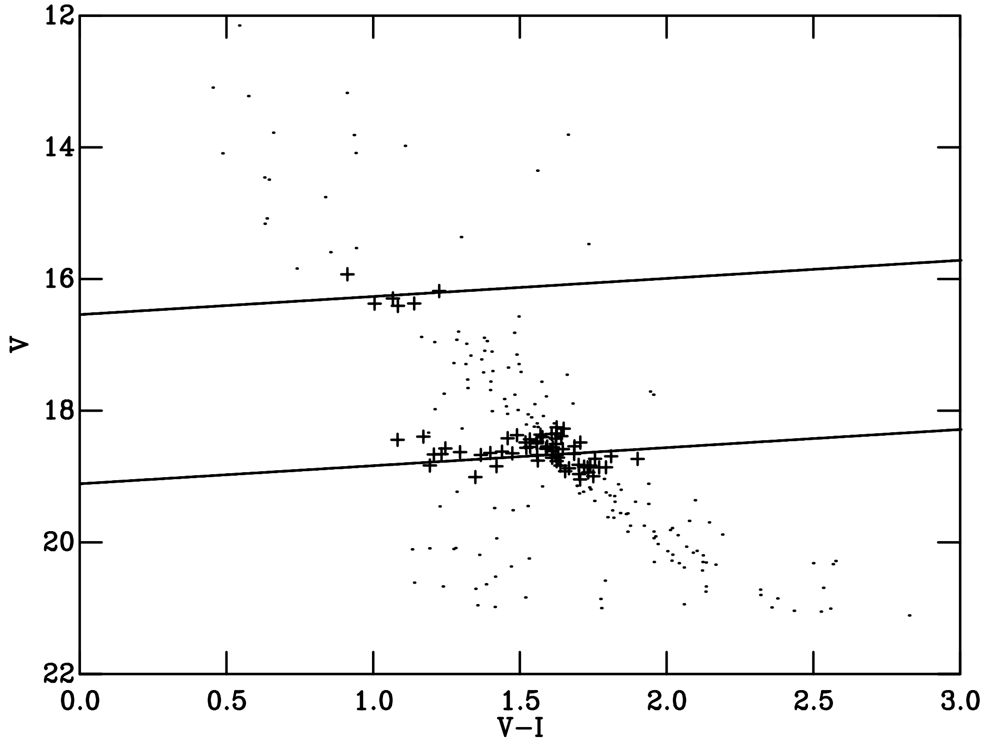

Figure 2.3: Placement of cuts along the line of the sequence of h Per variables. Stars falling within either of the cuts are represented as crosses. Identical cuts were used to analyse the general cluster membership.

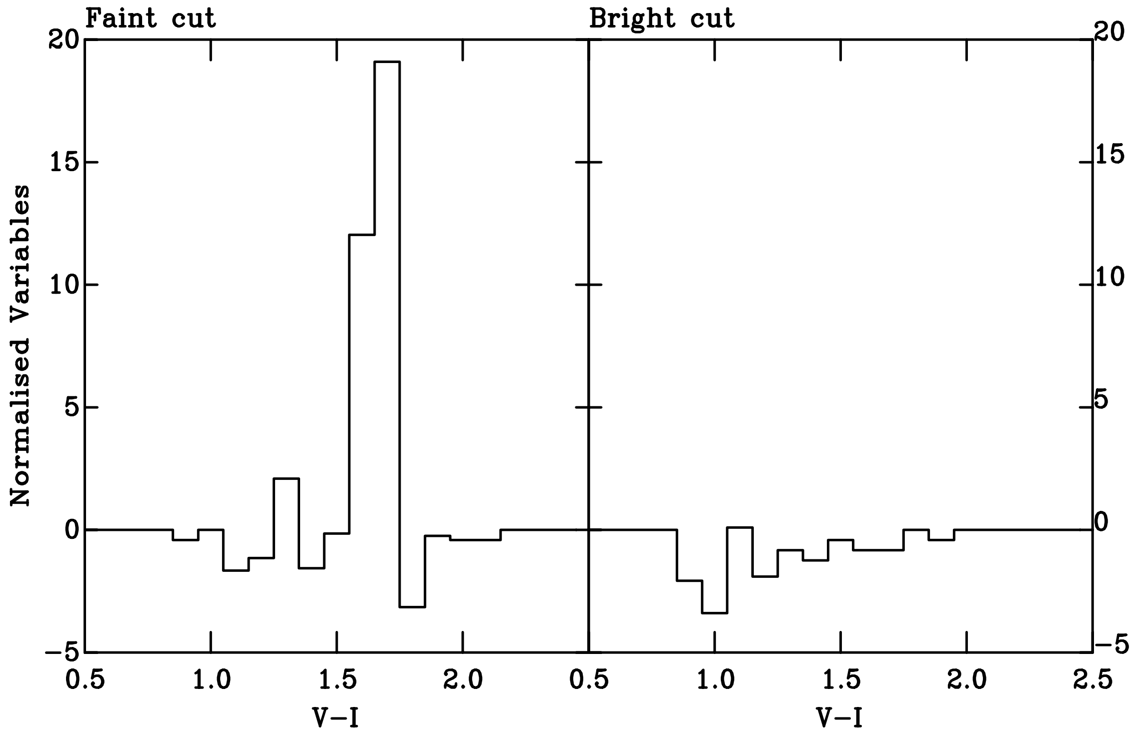

Figure 2.4: Histograms of the faint cut (*left*), made at 1.66 in V-I, and the bright cut (*right*), made at 1.0 binned along the line of the cut. Variable members falling with a 10' circle of the h Per centre are plotted, with the count of variable stars falling outside the cut normalised to the same area as the cut and then subtracted. The histograms thus measure the number of variables at each cut, after accounting for background variable contamination present in the field.

magnitude (lower) cut, the variables clearly trace the sequence at 1.6–1.7 in colour.

In the brighter magnitude cut (Fig. 2.4, right hand panel) however the situation is very different. There are relatively few variables present either in the background or in the sequence. In fact, the number of variables is consistent with a scenario in which there are no variables present in the sequence. The main point to take away from this is that although it seems possible, when looking at the CMD to trace the variable sequence at bright magnitudes by eye, closer investigation indicates that the number of variables is not statistically significant. There is therefore a genuine change in the number density of cluster variables, with variables dying out rapidly as we move to higher brightnesses. This is in marked contrast to the general cluster behaviour, which indicates the presence of a clear main sequence at these bright magnitudes. The implication is that beyond the R-C gap, the natural evolution of a previously variable young star leads to the complete suppression of that variability.

2.7 PMS star magnetic field topology and the R-C gap

Stolte et al. (2004) and Mayne et al. (2007) find evidence for a transition region between pre-main-sequence and main-sequence stars, which Mayne et al. (2007) term the radiative-convective (R-C) gap. The distinct change in the variability of the stars at a position that corresponds closely to the position of the R-C gap identified in Mayne et al. (2007), which is centred for h Per at V-I=1.33 (V=17.66), supports the conclusion that the mechanisms giving rise to the variability are a function of convective processes in the star, and are curtailed by the development of a radiative core. The obvious implication is that the stars below the gap have large-scale spot structures, and thus large-scale magnetic fields, whilst above the gap the spots are either absent, or sufficiently uniform in their distribution that they fail to yield a modulation. Thus these observations are consistent with the idea that whilst fully convective stars have large-scale structure in their magnetic fields, once the stars develop a radiative core the spots become more evenly distributed over the surface.

If this interpretation is correct, then we would expect to see the same change in variability in other clusters that show the presence of an R-C gap. Crucially, this will be at differing masses, depending on the age of cluster. Additionally, variability analysis should act as a tracer for the R-C gap. This has important consequences as it can be used as a distance-independent age estimator (Mayne et al., 2007).

In Saunders et al. (2008), Naylor explores the implications of applying a common theory based on changes in the stellar core to the problem of both the observed radiative-convective gap for PMS stars, and the so-called ‘period gap’ observed in studies of cataclysmic variables (CVs). For completeness, the argument presented there is briefly outlined here as follows. Cataclysmic variables are binary systems in which a higher mass white dwarf pulls mass from a less massive companion. Due to tidal effects, angular momentum is lost from the binary orbit as the mass transfer proceeds. Ultimately, the smaller star reaches a point where there is no longer enough mass present to allow radiative energy transfer to proceed efficiently, and the core undergoes a structural change to a convective mode of energy transport. This leads to a change in magnetic field topology that causes the ongoing mass transfer process to halt, and this is manifested as a range of CV periods that are not commonly observed.

The spectral type at which this angular momentum change for CVs takes place is around M4. In the work presented here however the radiative-convective gap is found to occur at K0-G8 in spectral type. This implies that a common theory cannot exist for both PMS stars and CVs that relies on internal structural changes to alter the field topology and hence create observable period changes.

Since the full scientific implications of this line of argument are tangential to the aims of this thesis, it will not be further pursued here. The interested reader is directed to Saunders et al. (2008) for a full treatment.

2.8 Conclusions

In this chapter I have described an analysis of variability in the young cluster h Per. It was noted that variability declines sharply at brighter magnitudes in the region of the R-C gap. This result is consistent with a cluster membership for which there are no variable stars at magnitudes brighter than the R-C gap. This is also consistent with the idea that variability arises from cool spots caused by magnetic field generation in the convective pre-main-sequence stars, and hence is not observed once a star reaches the radiative phase. Because the spectral type at which the R-C gap occurs is a function of age, if the variability is indeed linked to the location of the R-C gap, then it should not happen at a single mass at all ages — it must also be a function of age. This could be tested by performing a variability study of another cluster at a different age. If the variability is simply a function of mass, then the variable drop-off will not coincide with the R-C gap.

An interesting avenue for this research is to identify the distribution of periods for the variable stars identified in this study, and compare these to the established literature. Additionally, an analysis of period sensitivities would shed light on the sensitivity of this dataset to particular periods. Analyses of similar datasets indicates that diurnal aliasing is an inevitable problem, making periods around 1 day difficult to detect (see, e.g. Littlefair et al., 2005). Additional aliasing problems arising from the detailed timings of the observations may also be present. Figure 2.5 demonstrates the problems apparent in this dataset. Aliasing is severe, and the peaks are relatively broad, indicating signal uncertainty. Although techniques for improving the signal quality after the fact can make some impact on this problem, it is clearly better to improve the original data by placing the observations in the right place to begin with, if this is possible. A discussion of period sensitivity, along with periods and lightcurves for these variable stars will be presented in Saunders et al. (2008). However, these issues are exactly the kind of problems that a robotic telescope network is ideally placed to mitigate. In the following chapter, this idea shall be explored in detail.

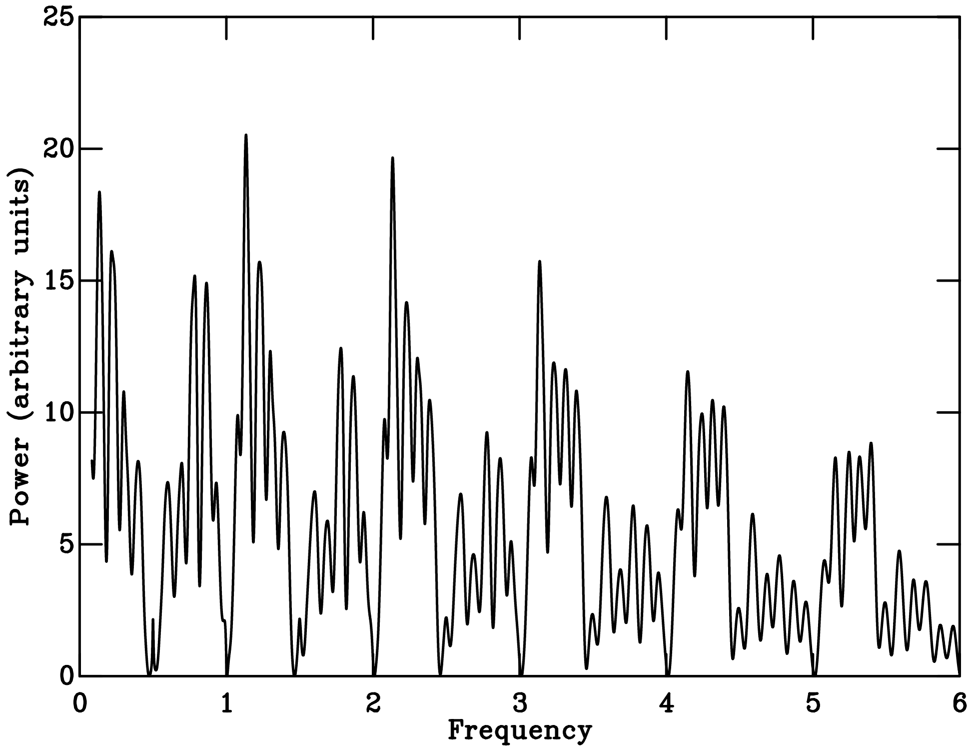

Figure 2.5: Periodogram of one of the variable stars from the h Per dataset, evaluated for periods ranging from 4hrs to 12 days. The severe aliasing is manifest.

Chapter 3

Optimal Placement of a Limited Number of Observations for Period Searches

3.1 Introduction

3.1.1 Motivation

Chapter 2 described an analysis of variability in the h Persei cluster. That work was based on 16 nights of contiguous observing from a single location (La Palma). This is typical for observations in the 'classical' mode. The long observing baseline was motivated by the desire to detect periods up to 8–12 days, while the continuous observing each night aimed to provide good enough phase coverage to unambiguously identify variability. Nevertheless, even with continuous coverage each night, the unavoidable, large, regular breaks in the time series caused by the diurnal sampling make periods of around one day difficult to detect in this dataset.

How well were these lightcurves sampled? The main characteristics of this sampling are dictated by the classical observing paradigm. Would an ideal set of observations be placed in the same way? As a purely theoretical question, could we have done better? Obviously, if observations could have been made continuously through the day as well as the night there would have been no phase coverage issues, and our sampling would be as dense as it could be. But this is not comparing like for like. The real question is more subtle. If we were to use the *same total amount* of observing time, then the new observation baseline is likely to be less than half the length of the original run[1] — leaving us insensitive to long periods. This suggests that the ideal set of observations is a compromise between dense, continuous coverage, and spacing observations farther apart in order to obtain a longer baseline.

In practice, this problem does not appear in the classical observing scenario because the limits are fixed and immutable. One can neither change the rising of the sun, nor the practical advantages (minimising the travelling time, cost, effort and inconvenience) of performing all the

[1] The actual fraction depends on the relative proportions of day/night at the observing location.

observations in one long set of nights. However, in the robotic paradigm, this is not the case. Observations can in principle be placed anywhere, and at any time, so the length of the baseline is theoretically unconstrained (although in practice it is likely to be bounded by the length of the observing semester, this is still much longer than a classically achievable baseline). If several robotic telescopes are available that are longitudinally distant with respect to each other, then an observer can effectively break the diurnal observing limitation.

The situation of limited numbers of observations arises in both robotic telescope and satellite observing. Typically target acquisition overheads, including slew time and CCD readout time, as well as scheduler requirements are substantial enough that a greater total target exposure time (and hence signal-to-noise) can often in practice be achieved by limiting the number of observations. There is also the more general question: how many observations are required to adequately address the science question under consideration? Are typical observing runs oversampled? If so, how much time is really required? From an efficiency standpoint this is a pertinent question for telescope operators and time allocation committees (TACs), as well as observers.

It is the norm for robotic observing runs to be undersampled. In this chapter the basis for this statement is discussed, and the implications for a general period-searching strategy in such an undersampled regime are explored. A number of sampling strategies are examined and discussed. It is demonstrated that there is scope for significant optimisation of observation placement, and a detailed description of how to implement such an optimal sampling is presented. The discussion has relevance for general problems in time series astronomy, and is not limited to photometry. For example, Doppler imaging of surface features (see e.g. Collier Cameron, 2001, for a review) is a spectrographic method that could potentially benefit from the work. The technique of *geometric sampling* presented here represents the best solution to the undersampling problem known to the author. However, it should be noted that the analytical solution to the general question of what form an optimal uneven spacing should take, when the number of observations and the length of the observing run are free parameters, remains unanswered. The main results described in this chapter have been published as Saunders et al. (2006a) and Saunders et al. (2006b).

3.1.2 The undersampling problem

The Nyquist-Shannon sampling theorem (Nyquist, 1928; Shannon, 1949) states that

> Exact reconstruction of a continuous-time baseband signal from its samples is possible if the signal is bandlimited and the sampling frequency is greater than twice the signal bandwidth[2].

Baseband describes a signal that has the lower bound of its range of frequencies at 0 Hz. *Bandlimited* means the signal has some limiting maximum frequency above which there is no further power. Thus the theorem describes a signal bounded in frequency. For an arbitrary signal, one can explicitly assume frequency limits, discarding information outside of the range. In this case, the theorem describes the conditions for complete reconstruction of the signal within the specified

[2]This formulation of the theorem is taken from `http://en.wikipedia.org/wiki/Nyquist-Shannon_sampling_theorem`. For the formal proof, see Shannon (1949).

limits. Equivalently, the theorem states that the required sampling rate for perfect reconstruction of a signal, called the *Nyquist rate* ν_N, must be twice the bandwidth (highest frequency component) of the sampled signal.

The determination of the Nyquist rate assumes regular, evenly spaced sampling. It is simple and unambiguous for the case where one can sample evenly across an entire observation run. However, astronomical observations are typically irregularly placed in time, with both short time-scale 'jittering' (intervals between observations may not be precisely the same) mid-range gaps (e.g. diurnal holes in coverage) and long gaps (e.g. long periods between successive monitoring campaigns of a single target). Robotic observations are even more irregular, typically with much larger gaps between individual observations. The deviation from regular, ordered sampling is significant.

A simple example shows that even if observations could be evenly placed, there are many observing scenarios where the required observing frequency is unachievable. For example, an astronomer observing T Tauri stars can effectively define a baseband, bandlimited signal by using the astrophysics of the problem to limit the frequencies of interest. A minimum period to be sought can be roughly determined by considering break-up velocity arguments, while the maximum period is limited by the coherence time-scale of the periodic feature (for example, star spots which may slowly change configuration over time), and more practically, by the maximum feasible run-length. For an evenly sampled dataset, the input signal may be correctly reconstructed if the sampling frequency is at least twice the highest frequency ν_max to be detected. Assuming the astronomer wishes to see two cycles of any modulation then the lowest frequency detectable is given by $2/T$, where T is the duration of the total sampling interval (i.e. the run-length). The value of the required sampling frequency ν_N for equally spaced data can then be viewed as the Nyquist frequency of the dataset, given by

$$\nu_\mathrm{max} < N/2T = \nu_\mathrm{N}, \tag{3.1}$$

where N is the number of observations. Plugging some typical numbers in shows the problem: if we assume a minimum period of around 4 hours, and a typical maximum period of around 2 weeks, then sampling at the Nyquist rate requires 0.5 obs/hr, or 168 observations, evenly spaced. On the Liverpool Telescope, a typical observation of a single field of a young cluster such as h Per requires around 300 s (5 minutes) of integration time (Saunders et al., 2008). To observe a reasonable fraction of the cluster, at least 3 fields are required with an instrument such as RATCam, which has a 4.6 square arcminute field of view (Steele, 2001). Thus an evenly sampled proposal would require 42 hours of on-sky integration time, before overheads — a substantial amount of time. At the telescope, contention from other proposals, as well as target acquisition overheads and the need to cover the time evenly makes this a difficult goal, and it is possible that the total number of observations obtained will be significantly less than requested.

When there are too few datapoints to fulfill this sampling condition, the case becomes interesting. In this situation can we can sample in such a way that we can recover a signal of acceptable quality? Alternatively, we can reverse the problem: how many datapoints do we need in order to recover an acceptable signal? In particular, it is important that the dataset for period searching is equally sensitive to different periods, lest the fraction of detected signals at different periods be misrepresented. This is a key problem, and one that has not been addressed in previous work.

3.1.3 Previous work

In the field of time domain astronomy, there is a relative dearth of literature regarding the best way to sample data in time. In contrast, much attention has been focussed on the problem of signal extraction from fixed, finite datasets (e.g. Horne & Baliunas, 1986; Schwarzenberg-Czerny, 2003). This is perhaps unsurprising. The vast majority of such datasets are acquired in the classical situation of a single extended observing run, where an astronomer physically travels to a remote observing site, accumulating data continuously each night for the duration of the run. In this case the observer has relatively little choice about how to space the observations, and normally opts for the safe strategy of using all of the available time for continuous observing, acquiring a large total number of observations.

A popular tool for the analysis of time series is the periodogram (Schuster, 1898; Lomb, 1976), defined as the modulus-squared of the discrete Fourier transform of the time series, appropriately normalised (Scargle, 1982). It presents the relative strengths of the set of harmonic frequencies making up the measured signal, and provides a powerful way of identifying periodic signals in the data, and determining those periods. Scargle (1982) extended the statistics of the Fourier transform to the uneven sampling case, and provided a metric, the *false alarm probability*, to quantify the degree of uncertainty in any given peak in the periodogram. Horne & Baliunas (1986) pointed out the sensitivity of the false alarm probability to the choice of sampled frequencies, and derived an empirical formula for estimating the number of independent frequencies required to calculate the false alarm probability in the uneven sampling case. The false alarm probability thus applied is often considered the primary indicator of the fitness of peaks in the periodogram, and forms the justification for many claims of signal detection to be found in the literature.

One common phenomenon that can cause problems in the interpretation of the periodogram is *aliasing*. It refers to the presence of spurious power in the periodogram at frequencies other than that of the true signal, which obstructs correct or unambiguous identification of the true period.

There are two principal causes of aliasing. In its common astronomical context, aliasing refers to power at incorrect frequencies in the periodogram arising from the presence of significant gaps in the phase coverage of the signal. The aliasing represents other possible periods which due to a lack of data cannot be ruled out. They are thus the manifestation of cycle count ambiguities (Kurtz, 1983). For example, Naylor et al. (1988) discuss how the sampling pattern of variations in the continuum flux of an eclipsing dwarf nova left them unable to differentiate between true orbital modulation and a secular decline in flux. This was a phase coverage issue arising from the observational constraints imposed by the use of a satellite.

A second form of aliasing arises from the beating effect between the sampling frequency ν_S and any true signal frequency ν present in the data. The beating causes constructive peaks located at regular intervals in frequency space at $k\nu_S \pm \nu$ for all positive integers k. It follows that the maximum frequency that is guaranteed to be unique occurs when $k = 1$, so that

$$\begin{aligned}\nu_{\max} &= \nu_{\mathrm{S}} - \nu_{\max} \\ \nu_{\max} &= \frac{\nu_{\mathrm{S}}}{2}\end{aligned} \tag{3.2}$$

(Kurtz, 1983). This is the Nyquist rate, and it defines the highest frequency at which the periodogram can be uniquely calculated. Frequencies beyond this value cannot be considered meaningful, since the periodogram is symmetric about the Nyquist frequency and contains no further information.

A number of authors have attempted to define an effective minimum sampling criterion for the case of unevenly sampled data. Scargle (1982) seems to imply that the effective Nyquist rate is a function of the smallest sampled frequency ("Averaging the data increases the effective sampling interval Δt, thus decreasing the Nyquist frequency" Scargle, 1982). Roberts et al. (1987) mention that in the uneven case, frequencies smaller than $1/(2\Delta_{\mathrm{min}})$ can be recovered, where Δ_{min} is the smallest interval in the dataset, but nevertheless choose to restrict themselves to a maximum limiting frequency of $1/(2\Delta_{\mathrm{min}})$, presumably as a convenient practical value. Similarly, Horne & Baliunas (1986) and Press et al. (1992) assume as an upper limit the Nyquist rate corresponding to an average sampling of points over the total run. Eyer & Bartholdi (1999) pointed out that the periodogram of an unevenly sampled dataset is not mirror symmetric about the Nyquist frequency, implying the presence of useful information above this value. They examined the practical limiting frequencies that could be recovered from unevenly spaced data, and demonstrated the recovery of a period of 0.085 days from a dataset in which $\Delta_{\mathrm{min}} \approx 0.8$ days, a factor 10 smaller than predicted by the conservative limits in the literature. Koen (2006) extended this work, and provided an explicit formula for determining the effective limiting Nyquist criterion. A remarkable example is provided of 53 unevenly spaced observations acquired over the course of 6.8 hours. The 'average Nyquist rate' for this run is $1/462\,\mathrm{s}^{-1}$, implying that the minimum period that can be accurately reconstructed from this dataset is around 460 seconds. However, Koen (2006) demonstrates the recovery of a 5 second signal that is clearly visible in the periodogram.

These various authors have provided a framework for understanding the recovery of high frequency signals in classically undersampled datasets. Importantly, they demonstrate that by breaking the degeneracy between observations, uneven sampling makes meaningful reconstruction of signals below the classical Nyquist rate possible. However, the critical problem for variability surveys is not the question of maximum frequency sensitivity. For many practical observing situations this value is effectively limited only by the precision to which the observations are recorded (Koen, 2006). The real issue is *even* sensitivity across the period range of interest. Although very high frequencies may be detected, typical astronomical observing runs remain undersampled. In extremis, a dataset could claim sensitivity to very high frequencies but be blind at periods of a day, because of aliasing effects. It is therefore possible to have good high frequency sensitivity, but very uneven sensitivity across the full period range. The undersampling generates ambiguities in an astronomer's knowledge of the signal, while the choice of sampling changes the relative sensitivity of the periodogram to different frequencies. It is therefore desirable to seek the spacing that maximises the information about the frequencies in which an astronomer is interested.

Clearly some unevenness can be coped with and corrected for by simulation, if an observer is only interested in determining the *distribution* of periodicities. This works as long as at least some of the variable population in the region of poor sensitivity are detected. However, if an observer is interested in obtaining specific periods then this technique cannot be applied. If stars are being analysed individually, then not having sensitivity at a given period is gambling.

Deeming (1975) showed that the observed Fourier transform of a discrete dataset can be expressed as the convolution of the true Fourier transform $F_{\mathrm{N}}(\nu)$ with a spectral window $\delta_{\mathrm{N}}(\nu)$ that fully describes the interference effects between the sampled frequencies, such that

$$F_{\mathrm{N}}(\nu) = F(\nu) * \delta_{\mathrm{N}}(\nu), \tag{3.3}$$

where $F(\nu)$ is the complex Fourier transform of a function $f(t)$ and is defined as

$$F(\nu) = \int_{-\infty}^{+\infty} f(t)e^{i2\pi\nu t}dt, \tag{3.4}$$

and the spectral window is given by

$$\delta(\nu) = \sum_{k=1}^{N} e^{i2\pi\nu t_k}, \tag{3.5}$$

where N is the total number of observations and $\{t_k\}$ is the set of observation times. The window function is even, so only positive frequencies are evaluated. The physical interpretation of the window function can be seen by considering the Fourier transform of an infinite sine function (a delta function) convolved with the window function, leading to the observed Fourier transform. Adjusting the set of observation times $\{t_k\}$ alters the window function, and hence the Fourier transform. Deeming (1975) showed that aliasing effects in the observed Fourier transform arising from the window function could be mitigated by introducing irregularities into the sampling pattern. Deeming (1975) termed the spurious power that is added to the Fourier transform by the sampling function *near-frequency interference*. In the discussions that follow, I shall refer to this effect as *spectral leakage*, after Saunders et al. (2006b). By careful choice of sampling it is possible to reconstruct higher frequencies than the ν_{N} required for an adequate equivalent even sampling, and to acquire higher quality data over the range of periods of interest.

Deeming (1975) illustrated the behaviour of the window function through the use of a simple single-variable model for the $N = 25$ case, empirically deriving a form for the spacing based on visual inspection of the resulting spectral window function. Motivated by similar concerns, scheduling for the Hubble Space Telescope programme "The Cepheid Distance to NGC 1425" adopted a pseudo-geometric spacing as a way to maximise the uniformity of the acquired phase spacing, but no details of the implemented optimisation strategy were provided (Mould et al., 2000).

3.2 Metrics

In the analysis that follows, the method of Scargle (1982), as formulated by Horne & Baliunas (1986) has been used to construct periodograms from simulated time series data. Many other approaches to period detection exist, including phase dispersion minimisation (Stellingwerf, 1978), string-length methods (Dworetsky, 1983) and χ^2 folding (Horne et al., 1986). The periodogram was chosen for reasons of convenience, simplicity and familiarity.

The metrics presented here aim to explicitly formalise some of the major processes by which astronomers arrive at a judgement about the validity of a period in a given lightcurve. Each metric provides information about a specific aspect of the lightcurve. The metrics are simple and unambiguous, in order to aid interpretation of the results, and for the other reasons discussed here.

In the literature, judgements about lightcurve periodicity have a somewhat qualitative nature, and often possess a significant subjective element (Herbst et al., 2000, is a typical modern example). This is perhaps an inevitable consequence of the application of detailed expert knowledge to the evaluation of complex astronomical datasets. By encapsulating some of the key elements of the knowledge required to understand and interpret such datasets in the form of metrics, something of the decision-making process has been made explicit, a useful exercise in itself. The metrics can also be used to evaluate existing datasets or partial datasets, and have proved useful to classical observers seeking to determine how best to utilise a second run of data to improve phase coverage or period sensitivity.

The metrics also provide objective functions that can form the basis of a directed optimisation process. They can be used to unambiguously specify utility, and thus provide a way for a computer program to interpret lightcurves. In Chapter 4, the use of such metrics in the practical implementation of autonomous agents for the eSTAR project (Allan et al., 2004a,b) is discussed. By providing empirical measures by which the fitness for purpose of a lightcurve may be determined, it is possible for software agents to reason about the quality of a given time series, and use that information to plan future observing requests.

In the results that follow, the use of the well-known false alarm probability of Scargle (1982) has been avoided. The false alarm probability measures how likely a peak of a given strength is to have occurred by chance, and thus provides an insight into the degree of structure present in the data. However, it says nothing about the validity of the period. Each peak of a periodogram with a serious aliasing problem has a low false alarm probability. Simply quoting the false alarm probability without acknowledging the other peaks is thus deeply misleading.

Additionally, the accurate calculation of the false alarm probability is non-trivial for the uneven sampling case. Horne & Baliunas (1986) note that the standard formula they provide is accurate only in the even-sampling case, and significantly underestimates the false alarm probability for severely uneven sampling. This is a serious issue for highly organised, non-linear observation timings such as those that may be acquired through the use of satellites or robotic systems.

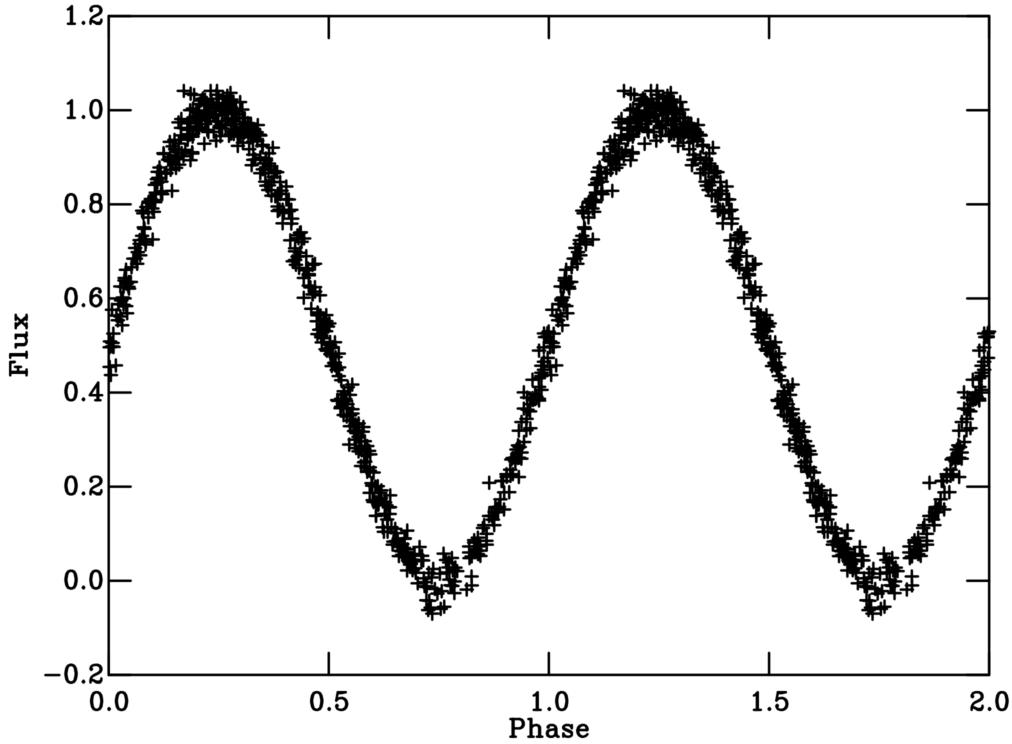

Figure 3.1: Folded lightcurve with $S = 2$, generated from the dataset of Littlefair et al. (2005).

3.2.1 Phase coverage uniformity (S)

In practice, many astronomers look at the folded lightcurve as a good test of periodicity. Consider Figures 3.1–3.3. Each figure is an artificial lightcurve, a regular sinusoid of fixed period sampled using the observation times of a real, classical observing run described in Littlefair et al. (2005). Figure 3.1 would be considered a reliable period. Figure 3.2 is reasonable, but by no means definitive. Figure 3.3 is simply unacceptable. The reason for these judgements is clear: Figure 3.1 covers the phase space well, while there is far too much phase information missing in Figure 3.3. In order to quantify this heuristic, we therefore need some kind of phase coverage metric.

The *normalised phase coverage metric*, S is defined to be the sum of the squares of the distances between consecutive points in phase space, normalised by the value of the sum for an ideal spacing (i.e. one in which all the observations are equally spread across the phase space). For an ideal spacing g_{I} of n observations we have

$$g_{\mathrm{I}} = \frac{1}{n}, \tag{3.6}$$

giving

$$S_{\mathrm{I}} = \sum_{\mathrm{j}=1}^{n} g_{\mathrm{I},j}^2 = n\left(\frac{1}{n}\right)^2 = \frac{1}{n}. \tag{3.7}$$

If the fractional part of an observation with timestamp x after folding on a period P is defined as $\phi = (x/P) - \mathrm{int}(x/P)$, then normalising the general case against the ideal spacing yields

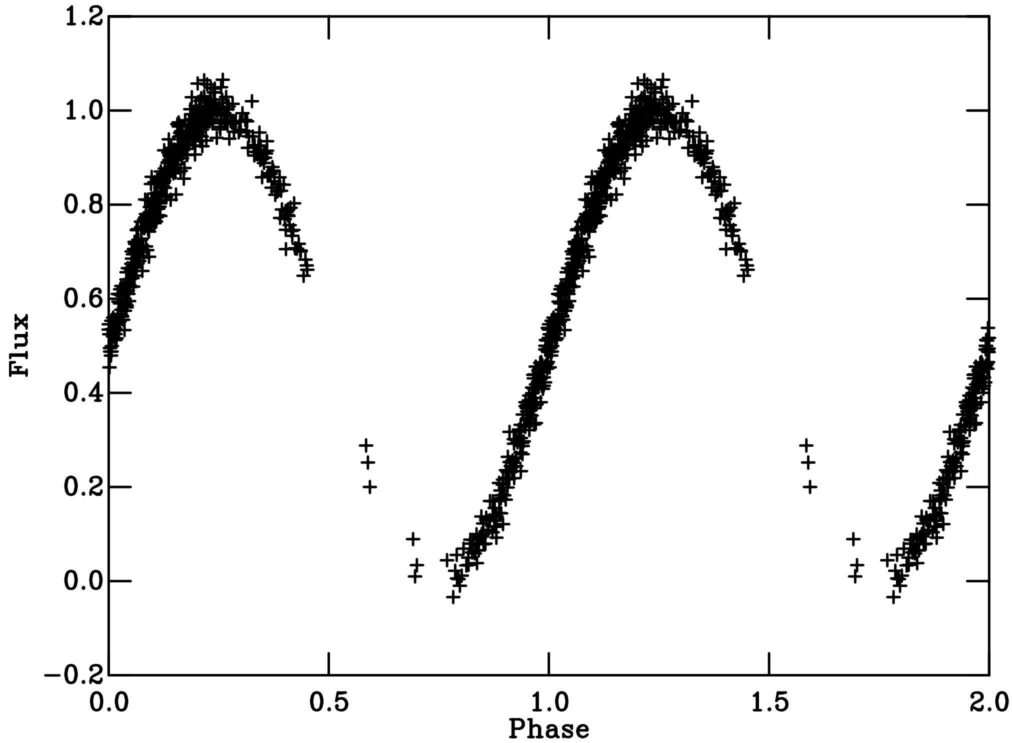

Figure 3.2: Folded lightcurve with $S = 20$, generated from the dataset of Littlefair et al. (2005).

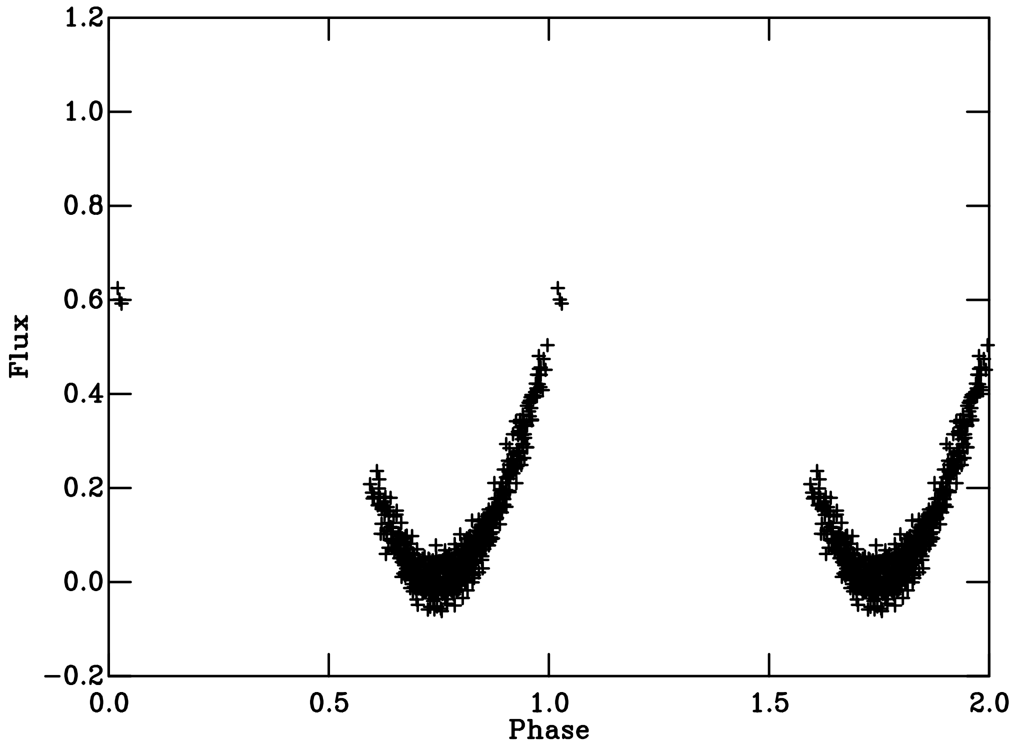

Figure 3.3: Folded lightcurve with $S = 200$, generated from the dataset of Littlefair et al. (2005).

$$S = \frac{S_{\mathrm{u}}}{S_{\mathrm{I}}} = n\,(\phi_n - 1 - \phi_1)^2 + n\sum_{\mathrm{j}=1}^{n}(\phi_{\mathrm{j}} - \phi_{\mathrm{j}-1})^2, \tag{3.8}$$

where n is the final observation in the folded sequence. The term $n\,(\phi_n - 1 - \phi_1)^2$ represents the wrapping property of the folding, by adding the contribution from the gap between the first and the last datapoints. This statistic is related to a variance, hence the choice of the symbol S.

The normalisation allows us to compare the phase coverage of datasets that have different numbers of observations. The metric provides an idea of how well a lightcurve has been sampled by the observations. To have confidence in a potential period it is important to know that the range of possible candidate periods has been minimised, i.e. that we have enough information to rule out other potential periods. As a result, lightcurves which possess better phase coverage tend to have much sharper peaks (i.e. much better frequency locality) in a periodogram.

Figures 3.1–3.3 illustrate how S acts to trace the overall phase coverage of lightcurves. A low S indicates good coverage, while a high S implies large phase gaps. Perfect phase coverage has the value unity.

The S metric as we have defined it here is simple and robust. However, there is a caveat. If we define the *relative period* to be the ratio of the period P to the run-length L, such that $P_{\mathrm{rel}} = P/L$, then for $P_{\mathrm{rel}} > 0.5$ the value of S is likely to be much closer to optimal, because there is no folding back of observation times, eliminating the possibility of redundant datapoints. It is normally good practice to sample over more than one signal cycle. Sampling over multiple cycles allows the repeatability of the signal to be tested. However, even the existence of multiple cycles may not be enough to avoid ambiguity. For example, a dataset sampled at 1.1, 2.2 and 3.3 cycles could once again confuse a temporal with a periodic modulation. S alone tells us nothing about how many cycles have passed in a given run-length, and gives a false indication of improvement out of proportion to the true value of the lightcurve for $P_{\mathrm{rel}} > 0.5$.

3.2.2 Interconnectivity (N)

The number of cycles over which a signal has been sampled is an important quantity, because it provides evidence of repeatability. Most astronomers would be loath to accept a period based on less than two cycles of data. An extension of this principle is that a folded lightcurve made up of consecutive points that are temporally distant in the original time series can be considered to have a more reliable period. Alternatively, the number of cycles can be used as a way of providing bounds for correlation timescales. How observations are distributed over cycles of the modulation can thus provide a measure of signal coherence.

We define the *interconnectivity*, N as the sum of the absolute difference of the integer part of the times of all adjacent observations, after dividing by a period P, such that for a set of n observations

$$N = |\mathrm{int}(f_1 - f_n)| + \sum_{\mathrm{j}=1}^{n} |\mathrm{int}(f_{\mathrm{j}} - f_{\mathrm{j}-1})|, \tag{3.9}$$

where $f_{\mathrm{j}} = x_{\mathrm{j}}/P$, x is the timestamp of the jth original observation, and P is the period under consideration. Note that f is different from the definition of ϕ used in Eqn. 3.8. f is the integer part of an observation after phase folding; ϕ is the fractional part after phase folding. The term $(f_1 - f_n)$ represents the wrapping property of the folding, by adding the contribution from the difference between the first and the last datapoints. The symbol N is employed for this metric, as it is a sum of integer values.

Together, S and N provide the tools to identify properties of the dataset such as period sensitivity and spacing efficiency. They have the great advantage that they are purely functions of time. They are independent of any properties of the actual observed data themselves, such as the signal shape or noise characteristics. They therefore cannot be biased by the lightcurve magnitudes, and are thus general tools for evaluating the sampling function of arbitrary datasets. This makes the application of these metrics straightforward and widely applicable. From a computational standpoint they are fast to calculate, scaling linearly with the number of observations.

3.2.3 Peak ratio (A)

The quality of the periodogram is assessed by calculating the *peak ratio*, A, which is defined as the ratio of the powers of the peak corresponding to the true period and the highest other peak in the periodogram. There are three regimes of interest for this metric. Figure 3.4 illustrates the 'alias' regime, where the metric indicates the extent to which power has been lost to other strong signal candidates. When aliases have been largely suppressed or when the overall level of noise in the periodogram is large, the metric instead describes the prominence of the peak relative to the spectral leakage, or to the background noise. These two situations are illustrated in Figures 3.5 and 3.6 respectively. For well-sampled datasets, alias signals can be suppressed to the level of the background noise. In this case A becomes a simple measure of signal strength. The relationship to the alias property or peak amplitude is the motivation for denoting this metric A.

The accuracy required for a given period measurement changes A. An observer would not be interested in all nearby peaks or sub-peaks, but the scale on which structure in the periodogram is important is highly dependent on the nature of the science programme. In the simulations that follow, periods within 10 per cent of the true period are considered to be adequate, in line with typical accuracy requirements for rotation rate histograms (Littlefair et al., 2005). It is assumed that the periods have an infinite coherence length, i.e. that the phase information does not change over time.

3.3 Simulation parameters

3.3.1 Assumptions

The metrics discussed above were used to investigate the behaviour of a range of simulated datasets. The general problem of ideal observation placement is extremely complex. Parameters that could affect the quality of an observed lightcurve include:

- The total number of observations.

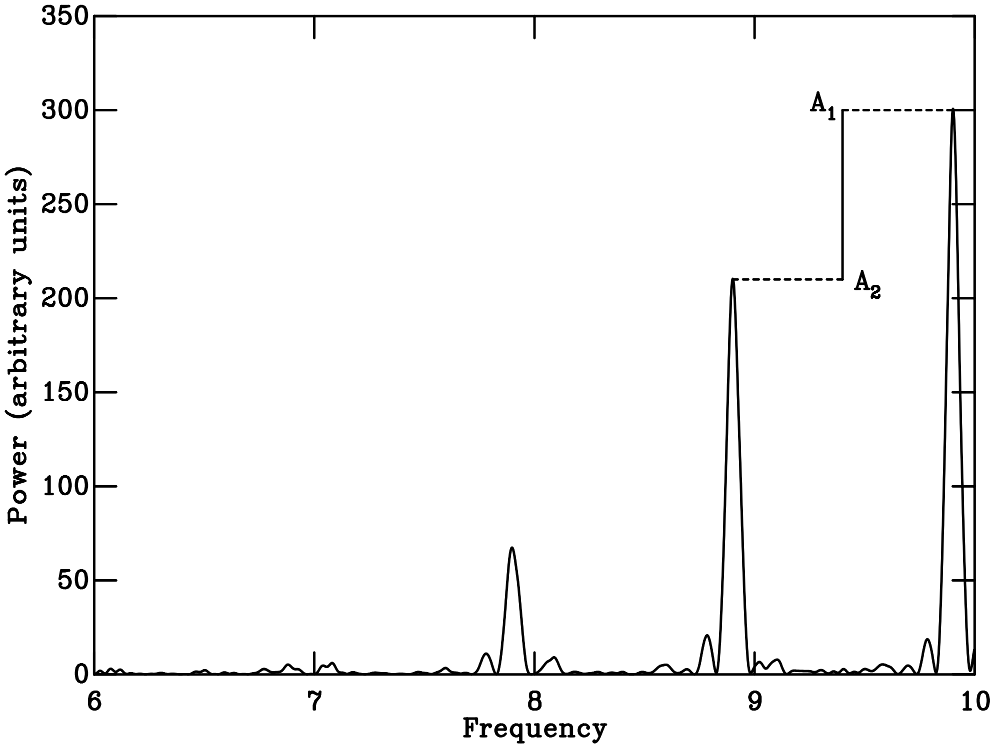

Figure 3.4: Periodogram illustrating how the A metric can be applied as a measure of relative alias strength. A_1 and A_2 are the amplitudes of the true signal peak (in this case, the highest peak) and the highest *other* peak in the transform. The ratio $\frac{A_1}{A_2}$ is the value of the A metric.

- The total length of the observing run.
- The choice of sampling strategy within a fixed window of observations.
- The period range of scientific interest (both the limits of the range, and the relative importance of different sub-ranges).
- The nature of the science performed. For example a survey of a large number of potential variables looking at general trends in variable population may not require period determination to be as precise as an extended campaign on a single object of interest.
- Observing constraints imposed by the telescope itself.
- The shape of the underlying signal. For example pulsating variables such as Cepheids have a distinctive saw-tooth shaped lightcurve. Observations made during the short, steep descent of the lightcurve may be more valuable than observations made during the slow rise.
- Noise, both correlated and uncorrelated. *Correlated noise* is noise which is a function of time, and thus is linked to the value of the noise at previous observations. This is effectively a second signal in the data that obscures the primary signal of interest. For example, a linear trend in the photometry may be due to some systematic effect within the detector, while a second periodic signal in the data might be due to a binary companion. Aperiodic variation, arising for example from changes in starspot configuration, is another source of correlated

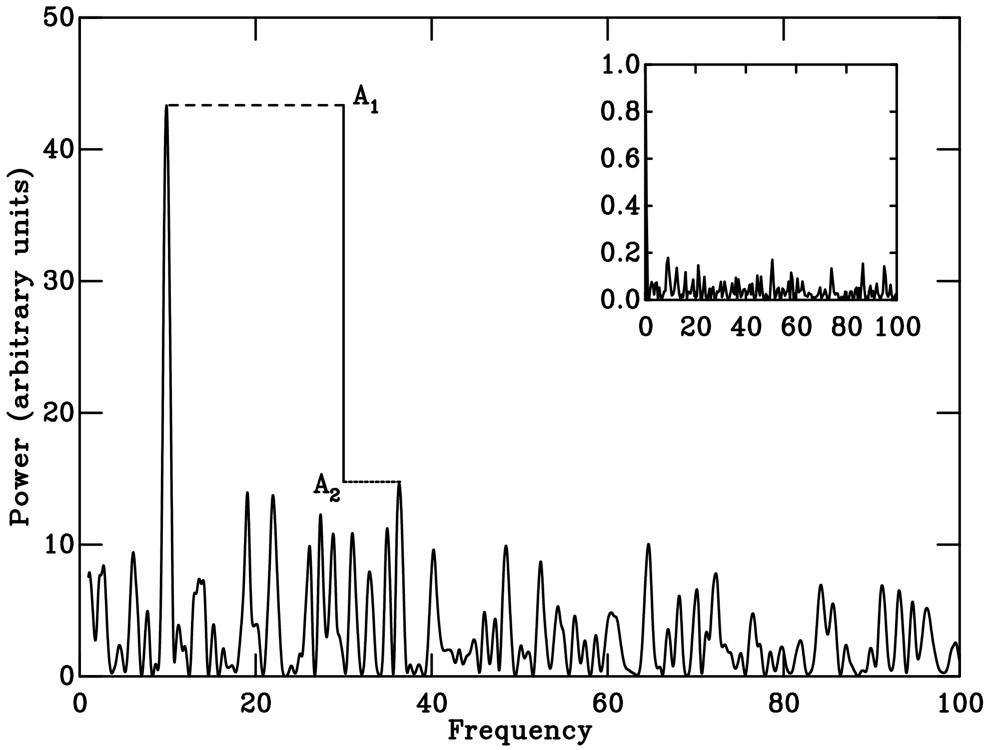

Figure 3.5: Periodogram illustrating how the A metric, defined as the ratio $\frac{A_1}{A_2}$, can represent the power at the true period relative to the spectral leakage. No noise has been added to this signal; interference effects are due to the (relatively poor) choice of sampling. The signal is a simple sinusoid with $P_{\rm rel} = 0.10163$, sampled with 100 datapoints using a geometric base of $x = 1.10$, randomly respaced (see Sec. 3.7). The inset gives the amplitude of the corresponding window function, as a function of frequency.

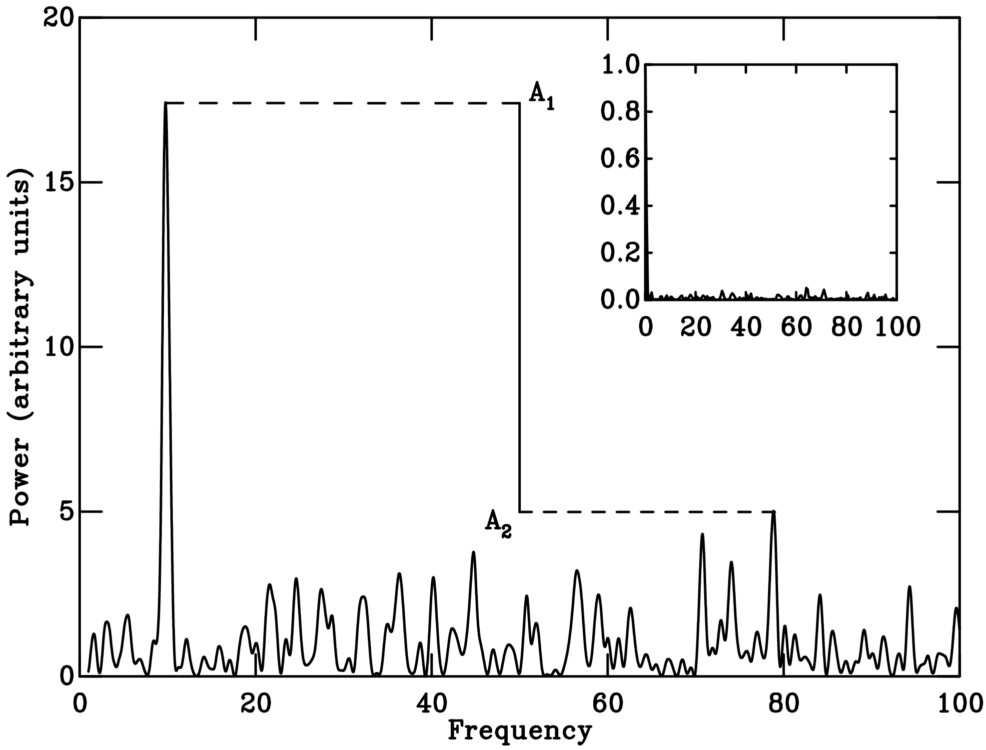

Figure 3.6: Periodogram illustrating how the A metric, defined as the ratio $\frac{A_1}{A_2}$, can represent the power at the true period relative to the overall level of noise present in the signal. The input signal is the same as for Fig. 3.5, but with the additional application of noise equal to half the amplitude of the signal . In this case, the geometric sampling base used was $x = 1.03$, to minimise the effects of spectral leakage (see Sec. 3.6). The inset gives the amplitude of the corresponding window function, as a function of frequency; note the relatively minor interference contributions.

noise, if the timescale for such changes is of the order of the run-length. *Uncorrelated noise* arises from the detector setup, the number of photons detected, measurement uncertainties and so on.

In the simulations that follow, timestamps in the range 0–1 were generated according to the sampling scheme under consideration. Where the value of A or the behaviour of the periodogram itself was considered, a range of sinusoidal signals of amplitude 1 with random phase were generated and the value of the signal evaluated at each sampling point. Realistic levels of uncorrelated noise were also applied (see Section 3.6.2 for the full details). S and N were calculated by evaluating the timestamps after phase folding at the known period.

3.3.2 Peak-finding algorithm

The determination of A was achieved by running a custom peak finding algorithm on the Fourier transform of each sampled signal. The peak-finder vertically processes the Fourier landscape, beginning at high amplitude and working down with successive sweeps. At each stage, a list of current peaks is updated by considering whether adjacent frequencies in the transform have amplitudes that fall within the current sweep. The highest value of the peak found so far is stored. The edge of a peak is defined by the imposed accuracy threshold (see Section 3.2.3), and individual frequencies may only be part of a single peak. In the case of the simulations that follow, this cutoff was set at 10%, so that any frequencies within 10% of the current highest peak with amplitudes above the current sweep level were evaluated as part of that peak. This strategy allowed the algorithm to avoid being trapped by sub-peaks and spurious noise, without compromising the peak resolution by looking at too broad a range of frequencies. Once the full transform was swept, the value of A was calculated by simply taking the ratio of the two highest peak values.

3.4 Linear placement

The simplest strategy for placing observations is even sampling. It is obvious that an evenly spaced sampling with enough observations will correctly sample a period range. Since the behaviour of such a regularly-spaced time series is well understood, it provides a useful baseline from which to investigate how the behaviour of the various metrics described in Section 3.2 respond to both adequately and undersampled datasets.

Figures 3.7–3.9 illustrate the variation of the three metrics A, S and N for three linearly sampled datasets, with 12, 100 and 1000 observations.

3.4.1 A

The undersampling problem is dramatically highlighted by the behaviour of the A metric. Figure 3.7 provides a clear illustration of the effects of undersampling in a regularly spaced sample set. The period range of interest spans two decades, from 0.01–1 in relative period. For Nyquist sampling an even spacing would correspond to 200 or more observations. For the two smaller datasets, this level of sampling is not achieved, leading to the presence of alias features in the

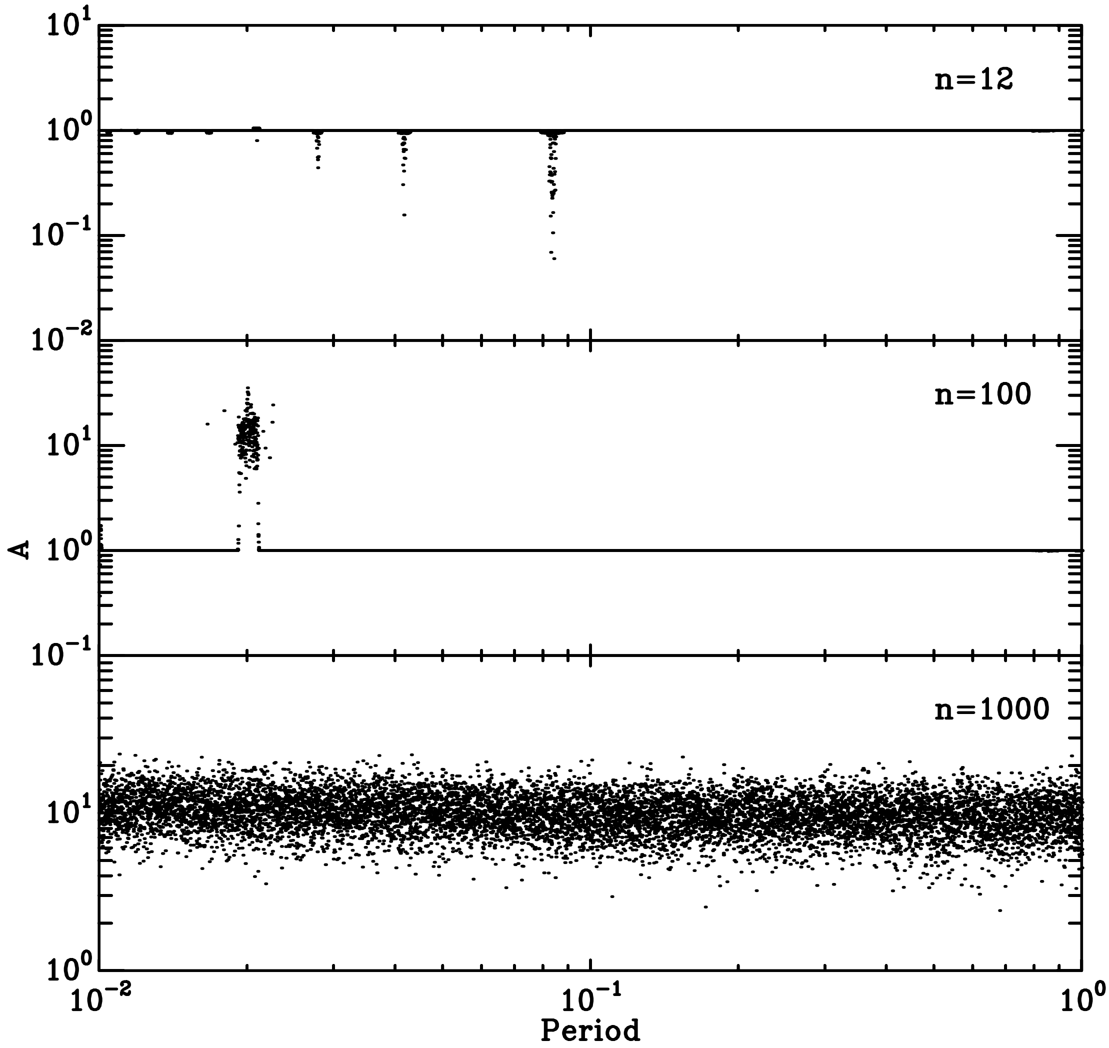

Figure 3.7: Variation of A with period for three linearly sampled datasets. The aliasing problem dominates the metric if n lies below the Nyquist sampling threshold of 200. A sharp spike occurs for the $n = 100$ dataset when the alias falls within the 10 per cent separation distance.

periodogram with power similar to the true period. This is reflected in the behaviour of A, which remains nearly constant at around 1 for $n = 12$ and $n = 200$, indicating the presence of a minimum of two very similar peaks in the periodogram. In contrast, for $n = 1000$ (Fig. 3.7, bottom panel) the sampling is adequate, and the value of A is relatively high. Even in the case of adequately sampled data, applying a small 'jitter' to the placement of individual observations can help suppress aliasing effects. In practice, some degree of jitter is automatically applied to real datasets, since it would be highly unusual for observations to be made at precisely the same time on every night of an observing run.

If the period range searched for each of our datasets is chosen such that the minimum period corresponds to the maximum recoverable frequency for that number of observations ($\nu_{\rm max} = \nu_{\rm N}/2$), then aliases do not arise. Figure 3.8 shows the value of A for the three datasets, when the period search space is thus reduced.

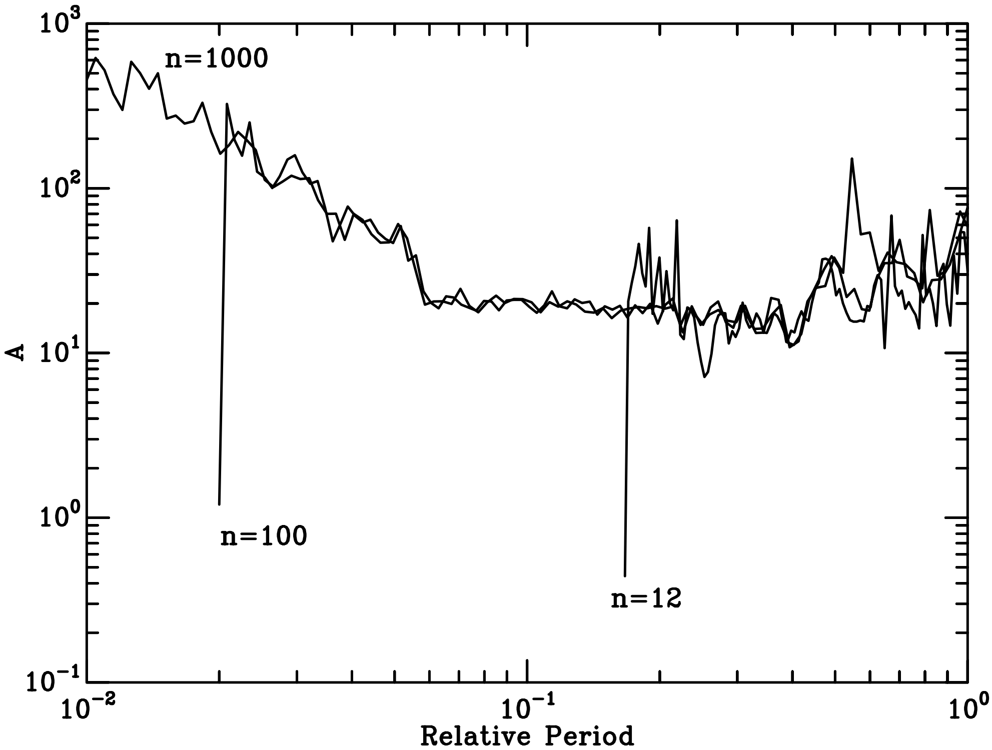

Figure 3.8: Variation of A with period for three linear observing regimes for which the maximum recoverable frequency has been set to the Nyquist frequency. As long as the frequency sampling regime lies above the Nyquist limit, the value of A is invariant with respect to the number of observations n.

This behaviour demonstrates how the false alarm probability is not useful as a metric here. Although the false alarm probability provides the probability that a peak could have been produced by chance, it says nothing about how many such peaks can exist as a result of aliasing.

3.4.2 S and N

A pure linear sampling, where observations are placed at fixed intervals in time, illustrates well the severe phase coverage issues that can arise from a poorly placed sample set. Figure 3.9 shows the effect of this sampling in terms of the S metric. Large spikes occur in the value of S at periodic intervals corresponding to multiples of the spacing value. These spikes arise because evenly spaced observations in time map back to the same points in phase space for a significant subset of possible periods. Such observations contribute nothing to the total information content of a folded lightcurve and are thus redundant from a phase coverage perspective. This is particularly bad for finding periods, because it implies an uneven sensitivity to periods of different lengths, which could introduce a systematic bias into a set of period determinations.

As the number of observations in a dataset is increased, the dataset becomes more stable, reducing the degree of structure. This arises because as n increases, the impact of any single observation on the overall phase coverage becomes less significant. See Figure 3.12 (discussed in section 3.5) for a clear demonstration of this.

Figure 3.10 (top panel) indicates how the interconnectivity N varies for a linear sampling with $n = 100$ datapoints. The structured nature of the period space is evident for relative periods of 0.5 or less (corresponding to true periods sampled by a total observing run over twice the period length) — the preferred regime for most observers. Sharp discontinuities occur when a favourable mapping interleaves many observations (producing a high N), while other mappings, very close in period space, lead to poor interleaving.

3.4.3 Conclusions

Linear undersampling demonstrates extreme behaviour for pathological cases with unfortunate periods. This makes it unsuitable for period searches with limited observations. Nevertheless, a linear sampling is the preferred sampling strategy when the dataset possesses enough observations to adequately sample the smallest period of interest. In practice, most of the problems of linear sampling described here can be overcome by 'jittering', the application of a small random component to the placement of each observation (Beutler, 1970). This is effectively the combination of a linear and a random sampling strategy. To understand how this changes the situation, we move in the next section to consider a purely random sampling strategy.

3.5 Random placement

While linear sampling provides a simple, unambiguous sampling rule when there are sufficient observations, robotic telescopes are in practice often observation-limited. Robotic telescopes can quickly become overhead-limited because of the fixed, finite time required to perform target ac-

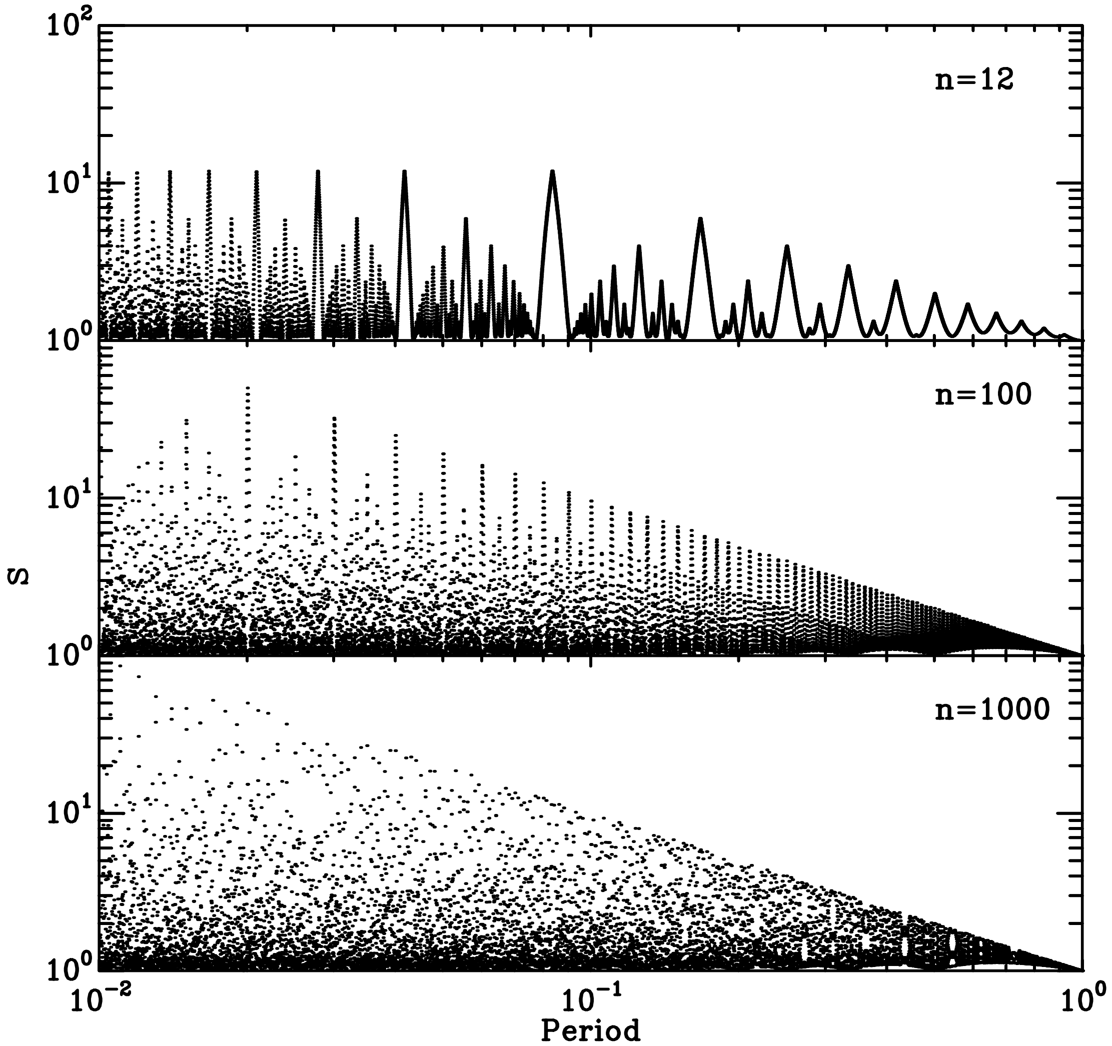

Figure 3.9: Variation of S with period for three linearly sampled datasets. Structure is evident at all dataset sizes, but is much more pronounced when the total number of observations is small.

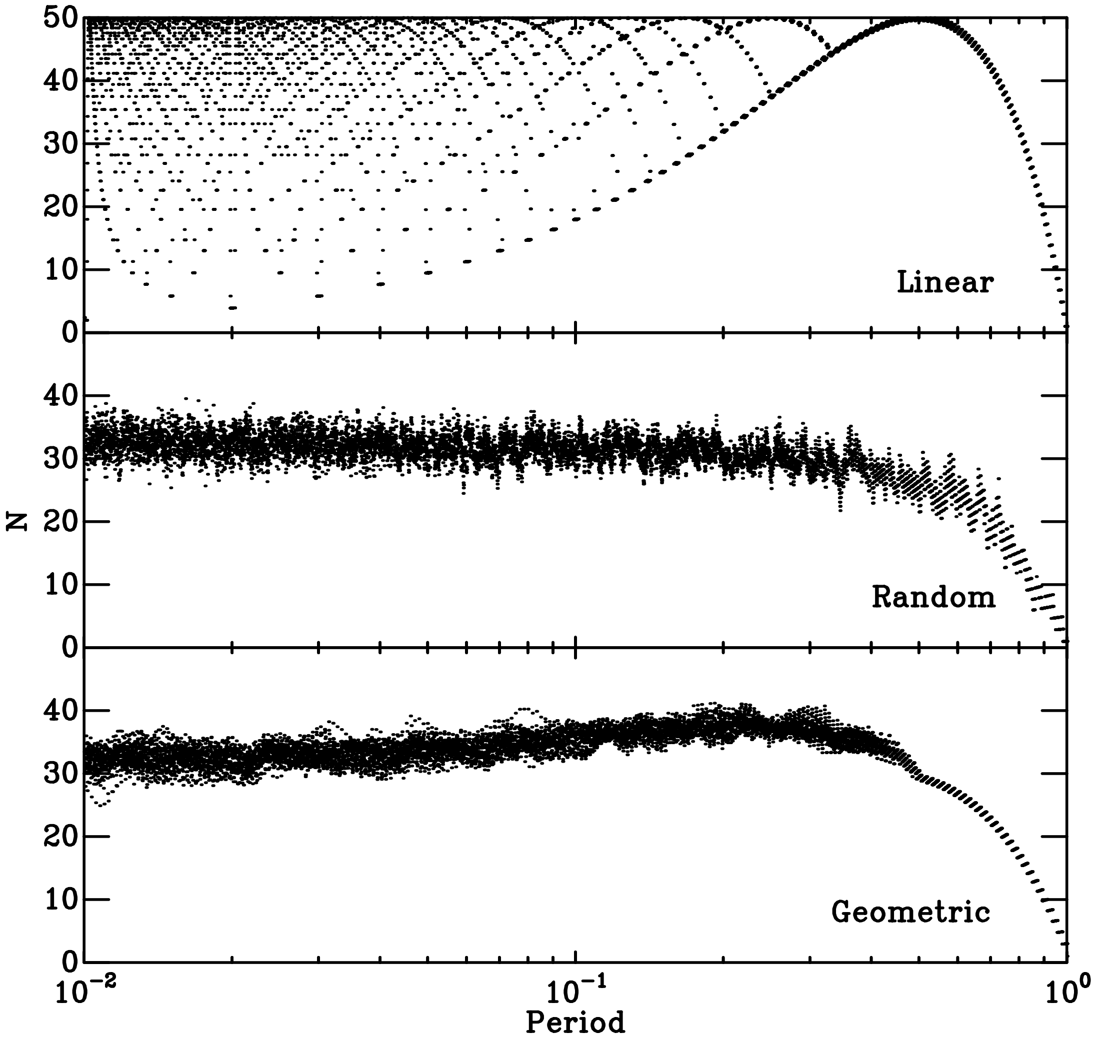

Figure 3.10: Comparing how N varies with period in the linear, the strictly random and the geometric case. For a pure linear sampling, (top) severe structuring is evident at lower periods. Outside the linear regime, N is less informative, because the relatively good phase folding means that the plot is dominated instead by the number of interleavings.

quisition. Space-based systems suffer from a similar problem. Once the overhead for each observation becomes a significant fraction of the total telescope time, a regime of sharply diminishing returns is entered, where taking additional observations is much more expensive than increasing exposure time. Random placement, an extreme form of jittering in the limit where the jitter length is equal to the total run-length, is an obvious alternative to a pure linear spacing. Each observation is placed at a randomly determined point somewhere along the run length, and no observation position depends on any other.

3.5.1 A

Figure 3.11 shows the variation of A with relative period for the three datasets. In all three cases, the value of A is greater than unity, substantially so for the $n = 100$ and $n = 1000$ datasets. This behaviour is in sharp contrast to the linear samplings of Figure 3.7, where the undersampled regimes remain largely static with respect to A, indicating the presence of strong alias peaks. Thus randomising the spacings has enabled coverage of the entire frequency range of interest.

The trade-off is increased background 'grass' across all frequencies of the periodogram. Comparing the $n = 100$ datasets of Figures 3.8 and 3.11 makes the point. When randomly sampled (Fig. 3.11, centre panel), A remains relatively constant at an approximate value of between 4 and 8, indicating a peak several times stronger than the background noise. In the linear case (Fig. 3.8) with the reduced period search space, the A value is much higher in the short period regime, and at its lowest point remains twice as strong as the equivalent random metric.

The background disturbance is present even when the random dataset has many more observations than would be required for Nyquist sampling, for example in the $n = 1000$ regime. There are two main causes of this noise. Because the observations are placed purely at random, it is quite possible for many of the gaps between observations to be much smaller than are required by our period range. These gaps are thus sampling frequencies outside of the range of interest, and do not contribute as much useful information to the sample set. Similarly, because the gaps are not precisely assigned, it is possible for particular areas of the frequency space to experience clustered sampling when pairs of observations fall, by chance, very near to one another in the frequency space. Both of these effects are manifestations of spectral leakage, and arise entirely from the choice of sampling.

3.5.2 S and N

Figure 3.12 shows the behaviour of S for a random distribution at each of the three dataset sizes.

In this random regime, S shows the stabilising effect of increasing the number of observations. As n increases, the amplitude of variation of S with period becomes much smaller. The extra observations essentially provide a kind of stability to the S metric by guaranteeing an average spacing between folded points which converges to the same value for large n, moderating the clustering effect described above in the discussion of A. This has important implications for period sensitivity. If the number of observations is too low, then the structure introduced in S implies that not all periods are equal: some periods will have intrinsically better phase coverage

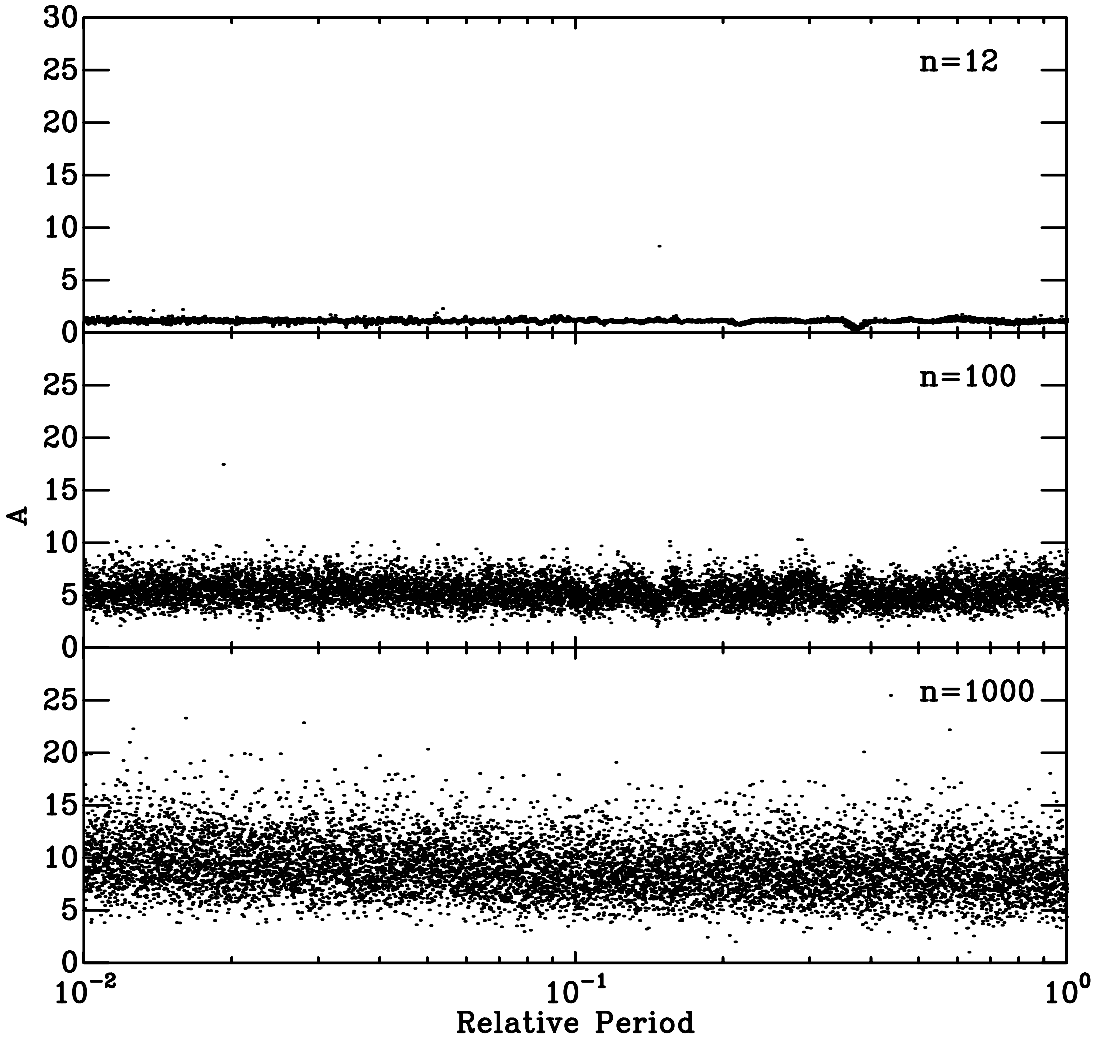

Figure 3.11: The variation of A against relative period, for a sinusoidal lightcurve randomly sampled at three different dataset sizes.

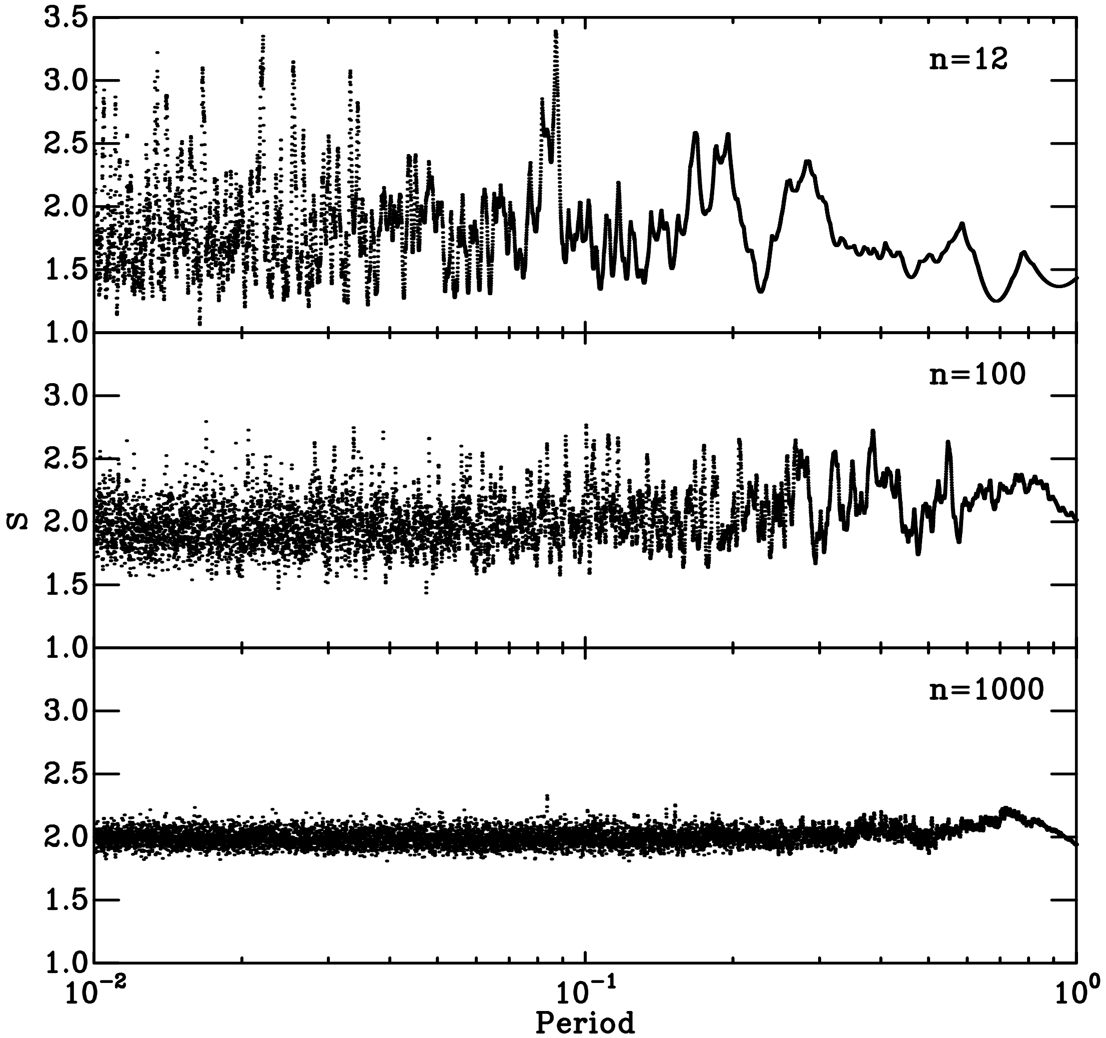

Figure 3.12: The variation of S against relative period, for 12, 100 and 1000 randomly selected observations of a sinusoidal lightcurve.

than others. When enough observations are made, random placement pushes S towards a value of 2, rather than the ideal best sampling of 1. This arises from the fact that S is essentially a measure of variance.

Figure 3.10 (centre panel) shows the variation of N for a randomly distributed dataset. The fine structure evident in the linear regime is no longer visible. This fine structure, the manifestation of the second term in Eqn. 3.9, is obscured by the large integer value of the interleaving provided by the first term.

3.5.3 Conclusions

By replacing a linear sampling with a randomised one, it is possible to cover frequencies below the Nyquist rate for even spacing. This coverage comes at the expense of increased background distortion in the transform (spectral leakage), the inevitable result of reducing the global signal coverage by spreading the observations over a greater period range. However, aliases arising from the undersampling are completely removed by moving away from the linear sampling regime. It is also seen that to some extent the periodogram is degraded because of redundant sampling, both by sampling frequencies outside the regime of interest, and oversampling frequencies within the regime as a result of chance. This suggests an improvement — explicitly choosing the size of the gaps between observations.

3.6 Geometric placement

The results of the random sampling experiments described in Section 3.5 show that the critical issue is the spacing of observations. This is obscured in the periodogram by the complicating presence of a signal, as well as any applied noise. Deeming (1975) considered the sampling problem in terms of the window function alone, deconvolved from the signal itself. This is an elegant and very general way to evaluate any arbitrary sampling, because it is independent of signal shape or amplitude.

A set of observing times specified by a single variable power law allow a wide range of gap spacings to be considered while limiting the vast search space of possible samplings. Another attractive feature of such a scheme is its intrinsic scale-invariance, a requirement for any truly general solution. We generate our observation times with a geometric distribution using the scheme

$$t_k = \frac{x^k - 1}{x^{(N-1)} - 1} T, \tag{3.10}$$

where t_k is the (time) value of the point in the series, x is the base of the series, k is the number of the current data point, N is the total number of observations and T is the duration of the total observing interval. This produces a geometrically-spaced distribution such that $0 \leq t_k \leq T$.

The parameter x, which shall be termed the *geometric base*, may be (almost) arbitrarily chosen. The generating function given by Eqn. 3.10 diverges at $x = 1$, which is the value corresponding to the limiting case of linear sampling. Values of $x > 1$ produce increasingly tightly packed sets of geometrically spaced points. Since the window function is in general complex, for comparison with Deeming (1975) we follow the convention of examining the normalised amplitude $A(\nu)$, defined as the square of the amplitude of the window function, normalised by the square of the number of datapoints, such that

$$A(\nu) = \frac{|\delta(\nu)|^2}{N^2}. \tag{3.11}$$

The *optimal base*, x_{opt} can then be defined as the value for which the overall degree of structure present in $A(\nu)$ is minimised. The root mean square (RMS) of $A(\nu)$ relative to the mean, given by

$$x_{\mathrm{rms}} = \sqrt{\frac{1}{n}\sum_{x=1}^{n}\left(A\left(\nu_i\right) - \bar{A}\left(\nu_i\right)\right)^2}, \tag{3.12}$$

where $\bar{A}$ is the mean amplitude, provides a straightforward metric for identifying structure. Minimising the RMS thus represents the desire of the astronomer to achieve relatively even sensitivity to periods. Note that this RMS is a function of the normalised amplitude $A(\nu)$, not a direct RMS of the window function amplitude $\delta(\nu)$.

In order to retain sensitivity to the higher frequency structure, $A(\nu)$ is evaluated for frequencies in the range $0.1 < \nu_{\mathrm{N}} \leq 5$ (in units of the Nyquist frequency for an equivalent even sampling). In practice, the range of the window function that needs to be considered will vary depending on the degree of undersampling of a given dataset (which is a function of the number of observations and the length of the dataset), and the maximum frequency to be detected. Since what matters is the frequency with respect to the total time interval, we define the dimensionless *relative frequency* as

$$\nu_{\mathrm{rel}} = \nu T. \tag{3.13}$$

Then the relative limiting frequency in units of the relative Nyquist frequency is simply $\nu_{\mathrm{max}}/\nu_{\mathrm{N}}$.

Although in principle the geometric base can be arbitrarily chosen, practical considerations provide a strong constraint. Consider a long observing run, with a baseline of one year. If observations are to be placed (for example) no closer than one second apart (although they could be spaced wider), then the maximum geometric base that provides spacings greater than this minimum is strongly constrained by the total number of observations, as illustrated in Figure 3.13. Thus the investigation is limited to the fraction of the power law space bounded by the curve.

3.6.1 Comparison with Deeming (1975)

Deeming (1975) generated 25 datapoints according to the formula

$$t \sim \begin{cases} k^{-1/\alpha} & (k = 1 \ldots 12) \\ (25 - k)^{-1/\alpha} & (k = 13 \ldots 24). \end{cases} \tag{3.14}$$

Figure 3.14 plots the window function for this distribution with $\alpha = 1.0$. This is a near-optimal value of α by visual inspection and RMS comparison. Figure 3.15 presents the window

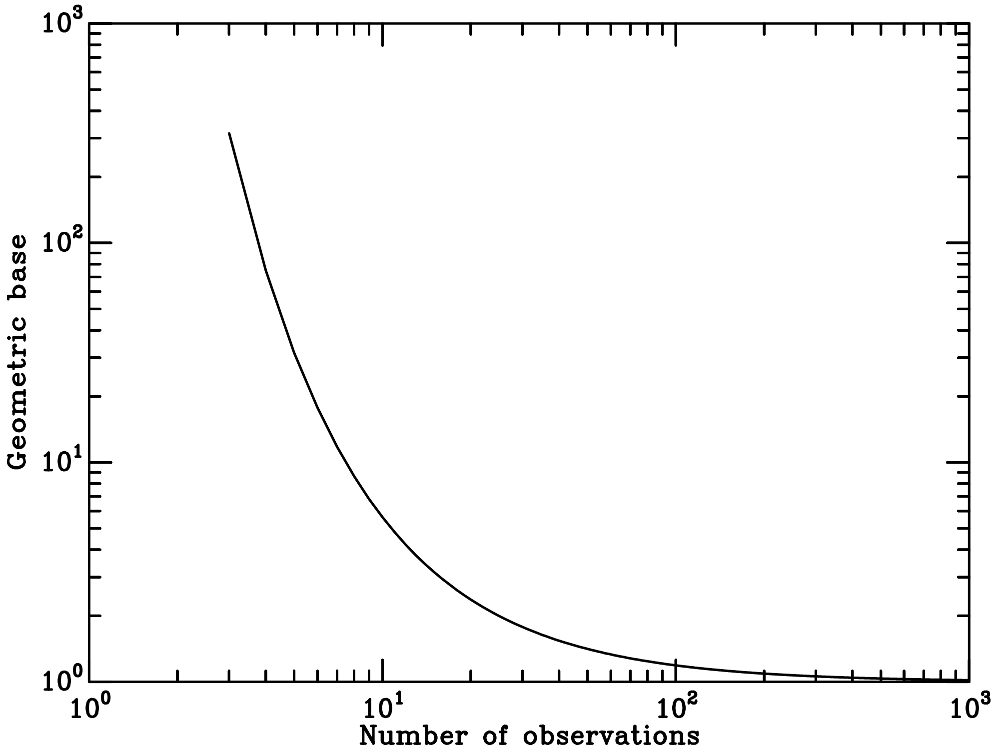

Figure 3.13: The maximum practical geometric base as a function of the total number of observations, for a given ratio of shortest period to run length.

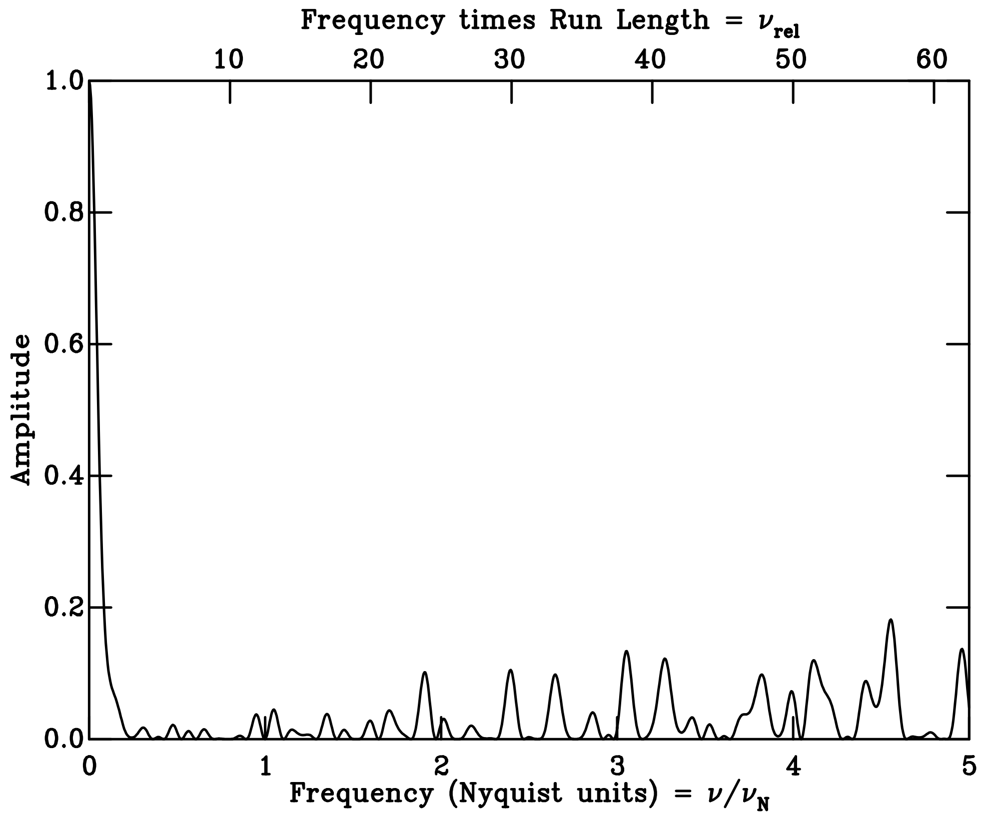

Figure 3.14: The window function (where the amplitude is given by $|\delta(\nu)|^2/N^2$) produced by Deeming's distribution, with a close to optimal value of $\alpha = 1.0$. The RMS is 0.0361.

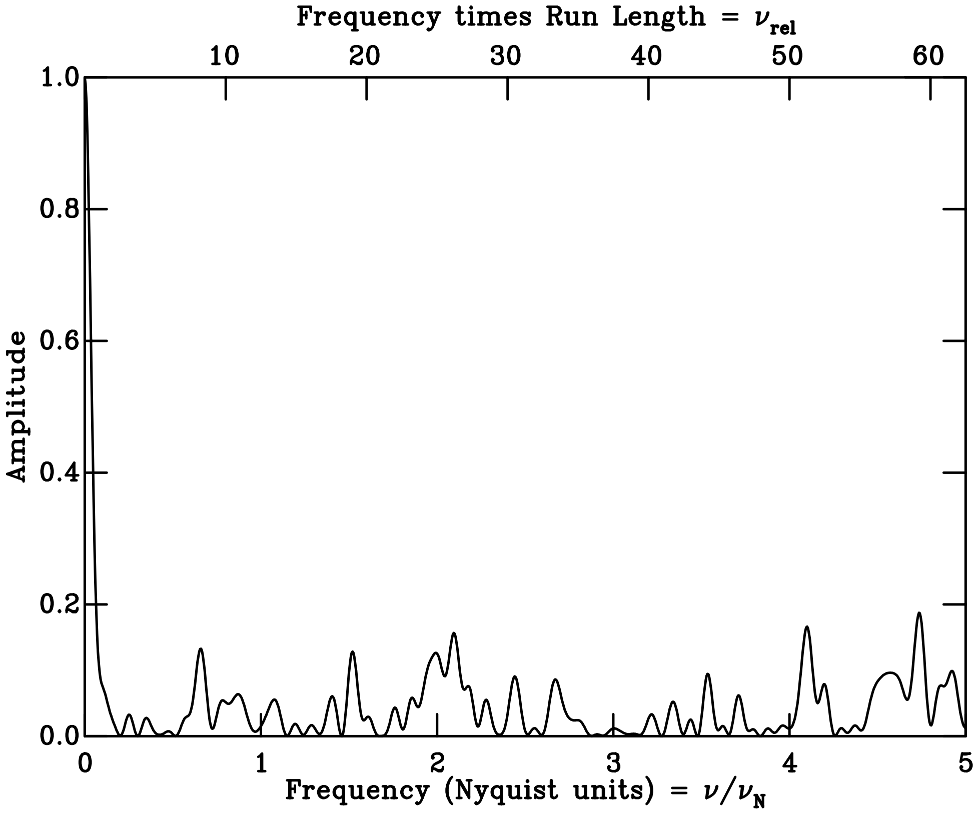

Figure 3.15: The window function for a typical random distribution of 25 data points. The RMS is 0.0386.

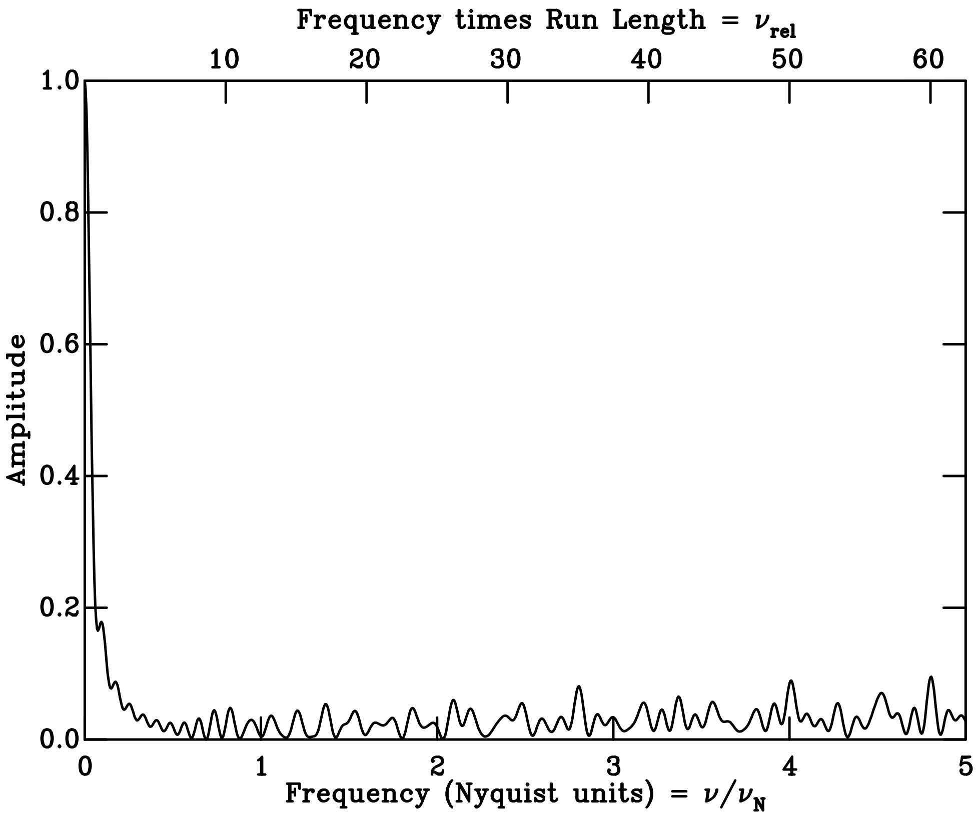

Figure 3.16: The window function for the optimum geometric sampling of 25 data points, for $x = 1.124$. The RMS is 0.0201.

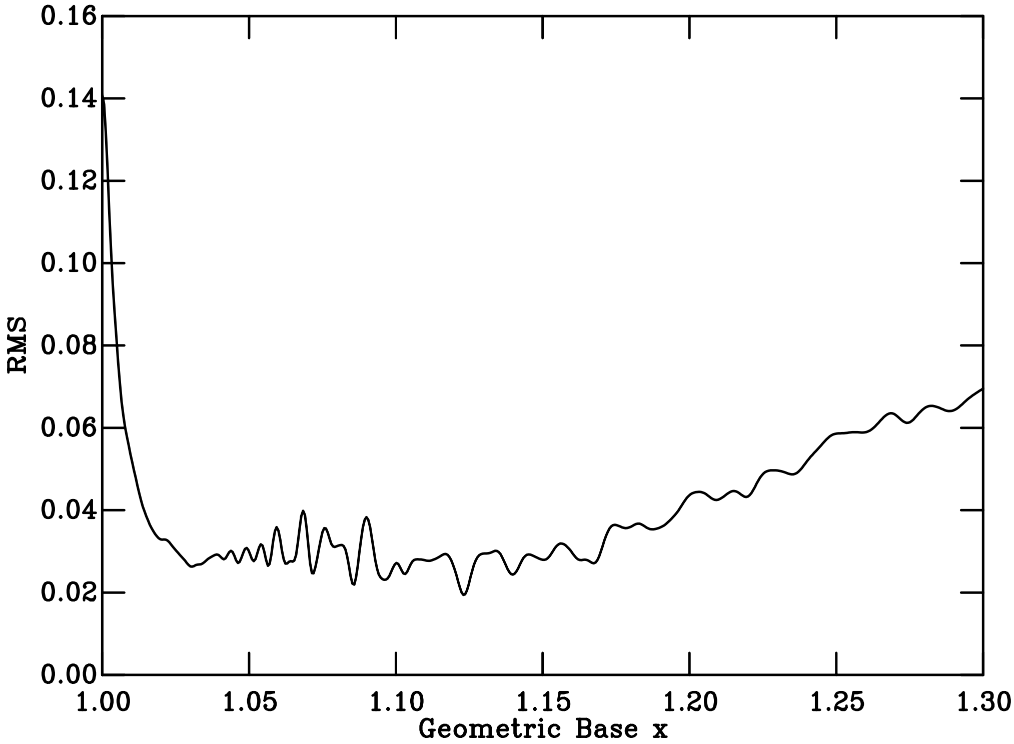

Figure 3.17: RMS as a function of geometric base, for $N = 25$. The optimum value is $x = 1.124$.

function for a randomly spaced set of 25 datapoints. The typical size of the interference structures at high frequency are comparable, as indicated by the similar RMS values of the two graphs. Figure 3.16 is the window function for an optimal geometric spacing, where the value of x was chosen according to the minimum of Figure 3.17. The geometric spacing exhibits much better high frequency suppression, with an RMS roughly half that of either of the other spacings, at the price of a slightly broader true peak (centred by construction at $\nu = 0$). This broadening is a consequence of the undersampling. Although the effective limit for frequency reconstruction can be pushed to higher frequencies, we sacrifice some of our knowledge of lower frequencies in order to do so. In particular, the precise frequency of the true signal is less clearly represented in the window function.

Note that the choice of geometric base is critical to the success of the sampling. Figure 3.18 shows how structure can be reintroduced into the window function by a poor choice of geometric base. This situation occurs because a sub-optimal geometric base essentially ‘wastes’ information, oversampling some frequencies and undersampling others. The pathological case of even sampling is a limiting example of this problem, where a large number of the sampled frequencies are massively oversampled, and therefore redundant. For datasets with more than 25 datapoints, the correct choice of geometric base is much more important. Figure 3.19 plots RMS as a function of geometric base for datasets with 25, 100 and 500 observations. The larger datasets possess much narrower minima ‘valleys’ than the $n = 25$ dataset. At these dataset sizes, any substantial deviation from the optimal base will lead to rapid deterioration of the window function.

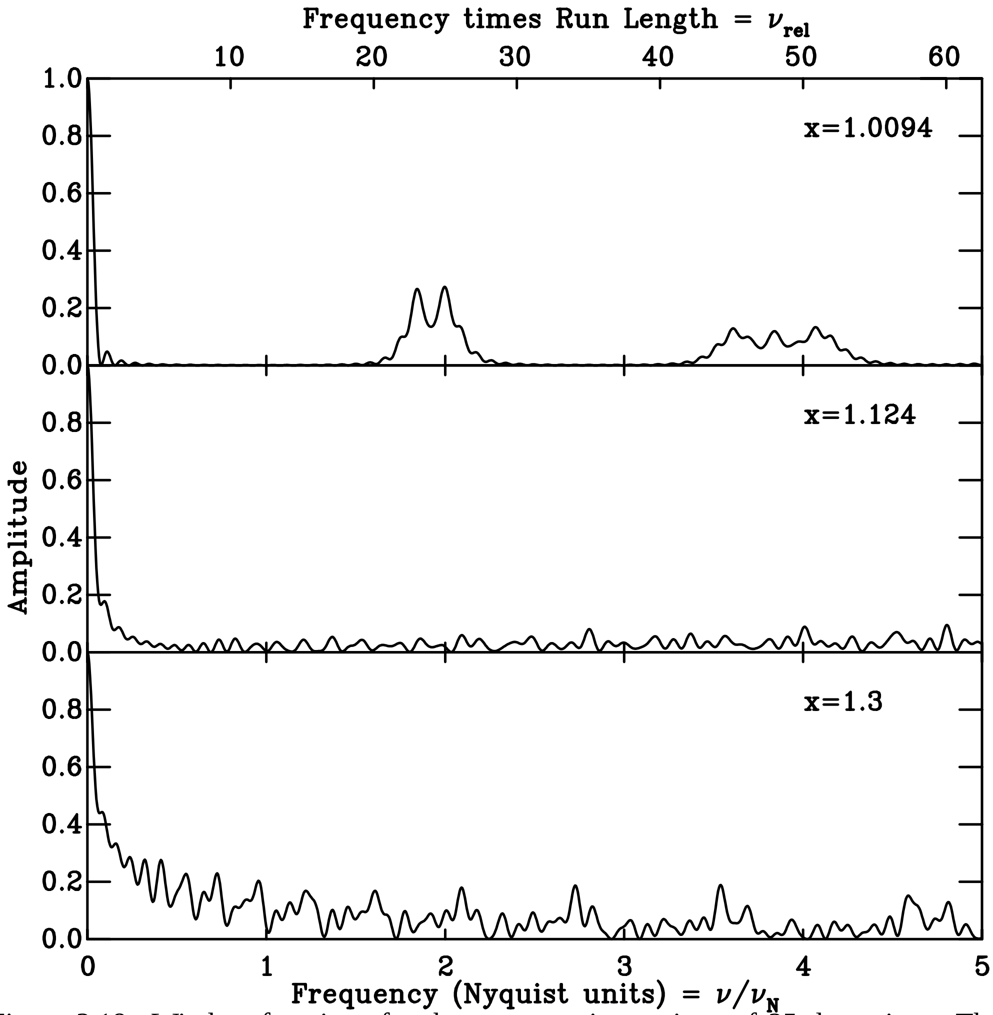

Figure 3.18: Window functions for three geometric spacings of 25 datapoints. The top panel illustrates aliasing effects at a near linear sampling of $x = 1.0094$, while at $x = 1.3$ (bottom panel) the effects of near frequency interference are apparent. The optimum geometric spacing for this number of datapoints is shown for comparison ($x = 1.124$, centre panel).

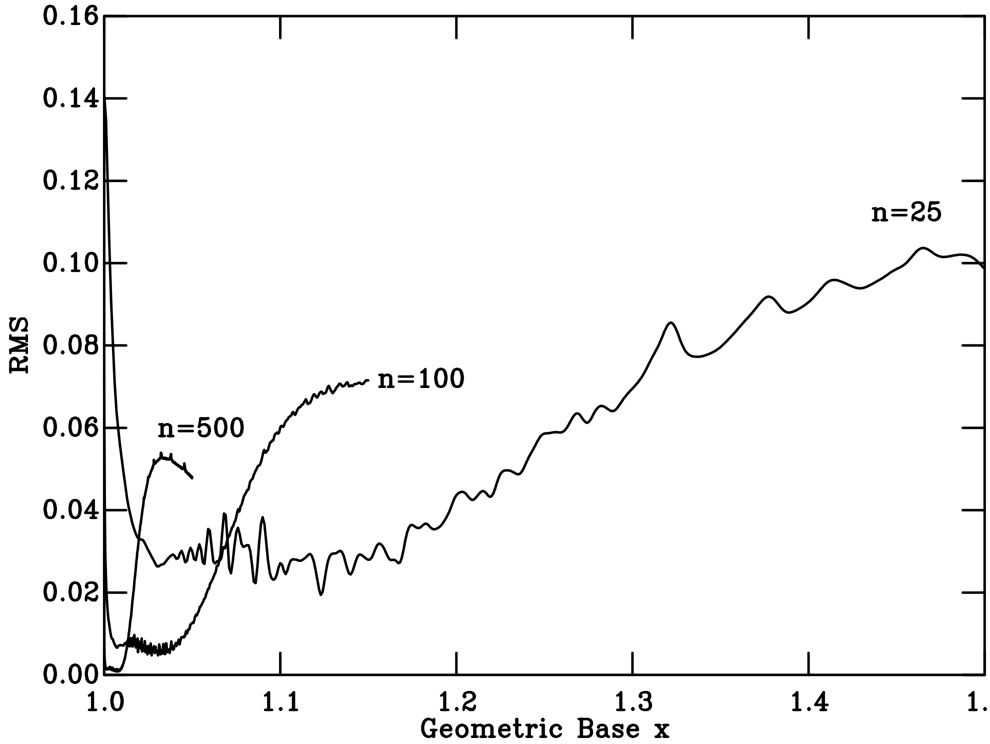

Figure 3.19: Comparison of the variation of RMS with geometric base, for three dataset sizes. When $n = 25$, a relatively broad range of bases lie close to the minimum RMS. For larger numbers of datapoints, the choice of optimal base is much more tightly constrained.

3.6.2 Noise

Whilst in the idealised case without noise a suitable geometric sampling is the correct sampling, the effects of noise on the result must be considered. If noise dominates the periodogram, then there may be little point in optimising in the manner described. A range of equally plausible 'optimal' geometric bases will exist, since the quality of the final observed Fourier transform is noise-limited rather than window function-limited.

When making actual observations, an observer must decide how to divide a fixed total amount of telescope time (exposure time plus overheads). This choice impacts signal-to-noise, since for most practical observing scenarios the signal-to-noise increases as the square root of the total exposure time. However, the window function has no noise dependency. The quality of a specific window function as a function of frequency is determined entirely by the choice of sampling time. This means that noise will only be of practical concern to the observing strategy if it is greater than the mean amplitude of the spectral leakage, the level of power added to the background frequencies by the choice of sampling.

To explore the practical effects of noise on the ability to recover a signal from a set of observations requires a return to the sinusoidal signal injection method described in the earlier results for linear and random sampling (see Sections 3.4 and 3.5). For a 100 observation dataset, noise is applied equal to one third of the full amplitude of the signal (Figure 3.20) and equal to the amplitude of the signal itself (Figure 3.22). This yields a signal-to-noise of 3.33 and 1 respectively.

It should be noted that this is not the signal-to-noise normally quoted by astronomers, where the time-varying signal is superimposed upon a constant flux. For comparison, in a typical observing scenario where the full amplitude of the signal modulation might make up around 10 per cent of the incident signal flux, the simulated noise described here would correspond to an observational combined signal-to-noise of around 33 and 10.

In Figure 3.20 the noise level is comparable to the spectral leakage, and we see the optimal base choice clearly picked out by the A metric. Lighter pixels indicate higher (and thus better) values of A. For comparison, Figure 3.21 gives the RMS of the window function as a function of geometric base for the same dataset. Figure 3.22 illustrates the degradation of the periodogram under the effects of extreme noise. Under these conditions, the choice of geometric base is much less important, because the amplitude of the noise is typically much greater than the amplitude of the spectral leakage.

The conclusion is that noise is not a contributing factor to the choice of optimal base for standard astronomical applications.

3.6.3 A

The existence of an optimal geometric base is generally true for any number of datapoints. However, there are some interesting qualitative differences. Figures 3.23 and 3.24 plot the three-dimensional space of possible power law base against relative period in terms of A for the 12 and 1000 observation datasets.

The critical feature at all dataset regimes is the presence of a bright vertical band of optimal bases, for which the value of A is maximised. The optimal base shows no variation with period because it is governed solely by the window function, which is itself independent of the signal. In all cases, optimal bases lying above this limit produce suboptimal samplings. The precise value of the optimal base depends on the number of observations. For the $n = 1000$ dataset of Figure 3.24, the lightcurve has been sampled with many more datapoints than the 200 uniformly spaced observations required to achieve Nyquist sampling. In this situation the result that linear sampling provides the best coverage of the period range is recovered.

If the number of observations is small, as in Fig. 3.23, then a degree of structure is introduced to the base-period plane, because the size of the gaps between observations is large enough to allow poor folding for some period ranges. This effect is simply a manifestation of the degree of coarseness of the sampling. The areas of poor phase coverage appear as long dark swathes that grow in width as the period of the signal approaches the run-length. As the number of observations in the geometric series is increased, these gaps shrink, although very fine structure can still be discerned even for the $n = 1000$ dataset.

The geometric base that maximises A for a given number of observations is that which generates a sampling distribution that best covers the desired frequency range. By placing observations geometrically, broader frequency coverage can be regained, at the expense of detailed coverage of narrower frequency bands. Without adequate broad frequency coverage, aliasing problems dominate the periodogram, as previously discussed for the linear case. Thus the ideal geometric base represents the balancing point between sharp period ambiguities and a generally poor phase

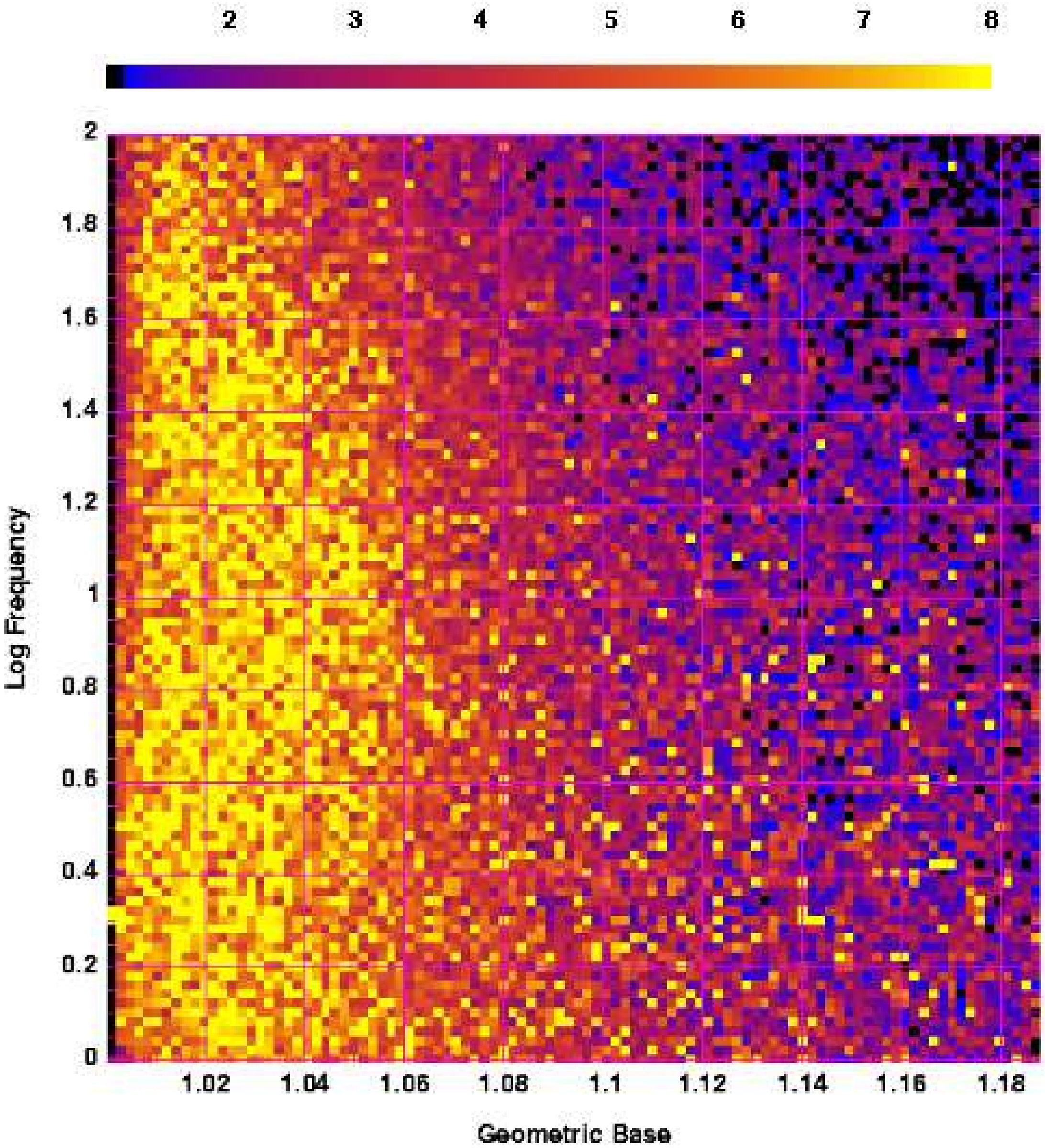

Figure 3.20: A for an $N = 100$ ordered geometric lightcurve with noise equal to 30% of the signal amplitude, as a function of geometric base and injected signal frequency.

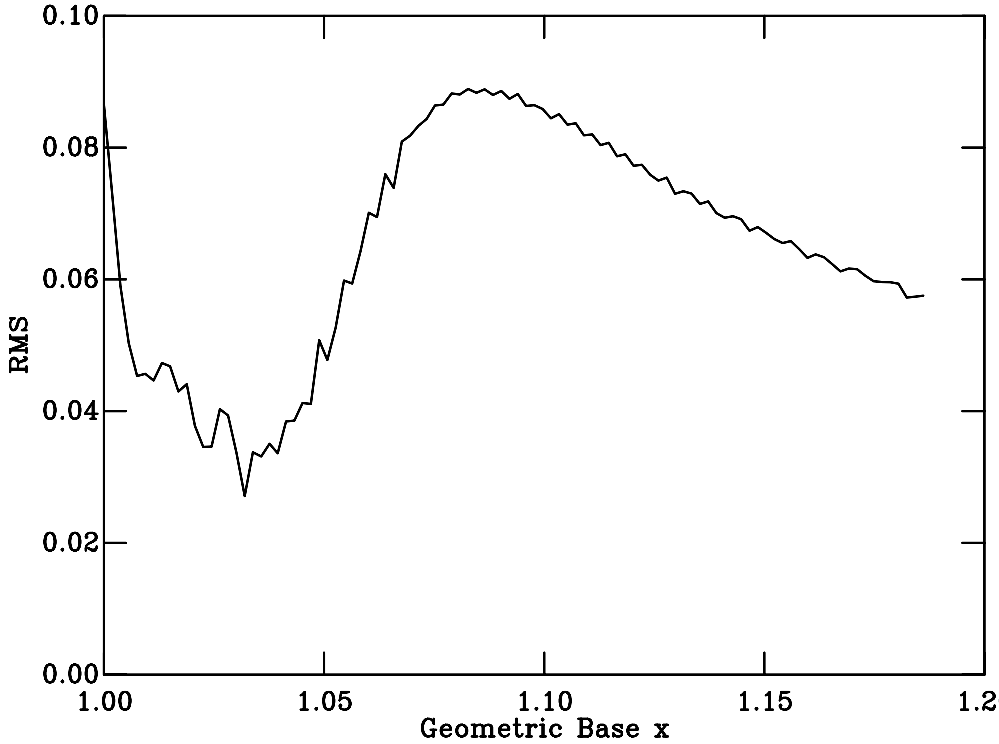

Figure 3.21: RMS of the window function, as a function of geometric base, for $N = 100$. The optimum value is $x = 1.032$.

coverage that produces power over a wide range of frequencies. It is the spacing at which the maximum value of the spectral leakage falls to the level of the background noise.

A higher geometric base corresponds to a greater clustering of data points towards one end of the observing run (for timestamps generated from Eqn. 3.10, this clustering occurs at the beginning of the run), which alters the frequency sampling in favour of smaller gaps. This is beneficial until the point at which the gaps become so small that a disproportionate number of very high frequencies have been sampled, at the expense of coverage at lower frequencies. Naively, one might expect this limit to occur when the size of the first gap is around half the width of the minimum period of interest. However it was found empirically that in fact the optimum choice of geometric base has a minimum sampling frequency somewhat smaller than this number. This is consistent with the descriptions in the literature discussed in Section 3.1.3, where it was shown that it is possible to accurately identify smaller periods than expected.

An intuitive way to visualise what is happening is as follows. Consider a sampling of only two points. This provides only one gap, and hence only one sampled frequency in the resulting Fourier transform. If a third point is added, however, it interacts with both of the previously placed points, producing a total of three frequencies. In fact, for n observations, the total number of frequencies F sampled is a triangular number (the additive analogue of the factorial) given by

$$F = (n-1) + (n-2) + \ldots + 1 \quad \text{(where } n \geq 2\text{)} \tag{3.15}$$

Thus if the smallest gap is exactly sufficient to sample the highest frequency of interest,

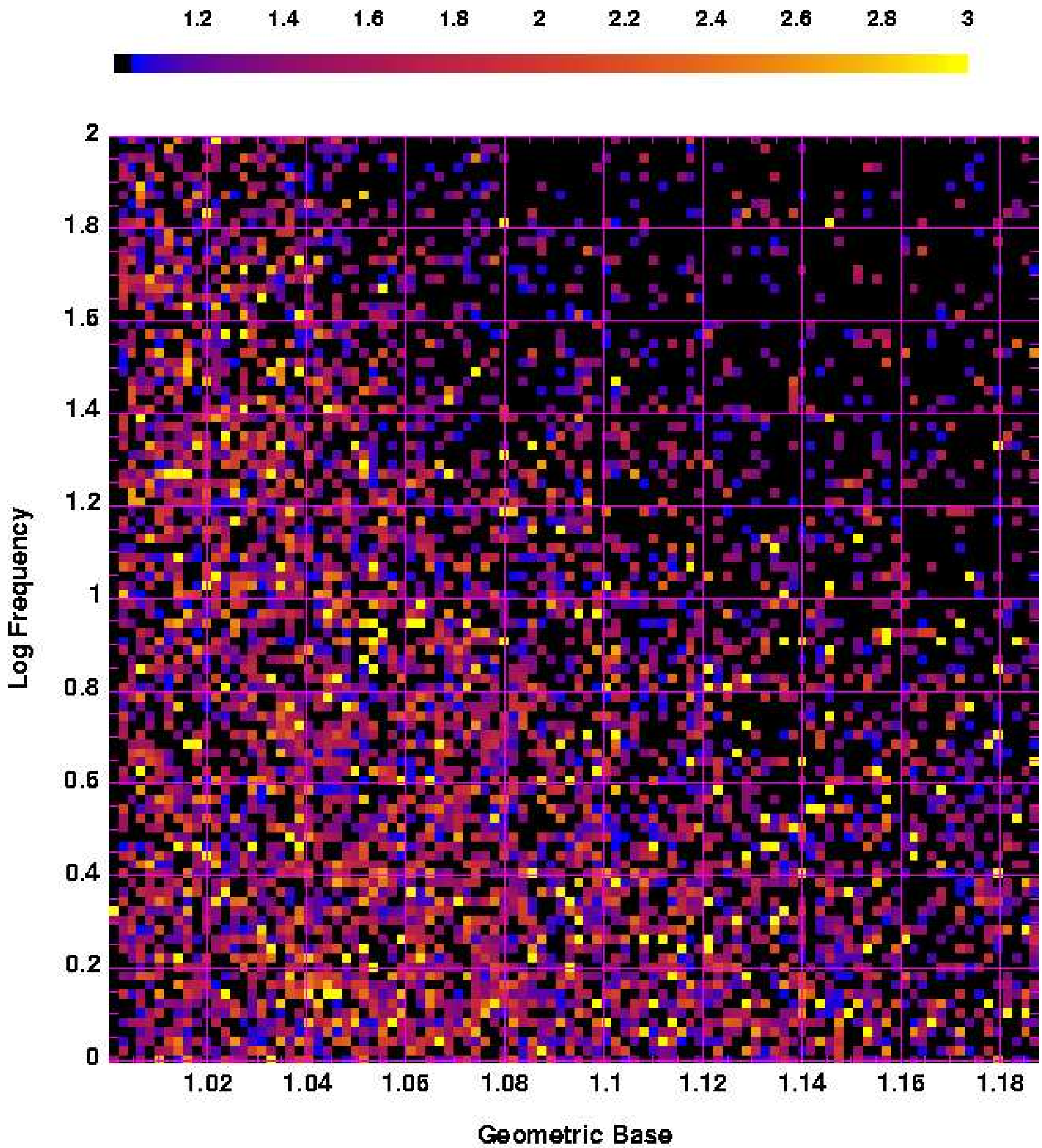

Figure 3.22: A for an $N = 100$ ordered geometric lightcurve with noise equal to 100% of the signal amplitude, as a function of geometric base and injected signal frequency. High noise dominates the periodogram, washing out structure in the plane.

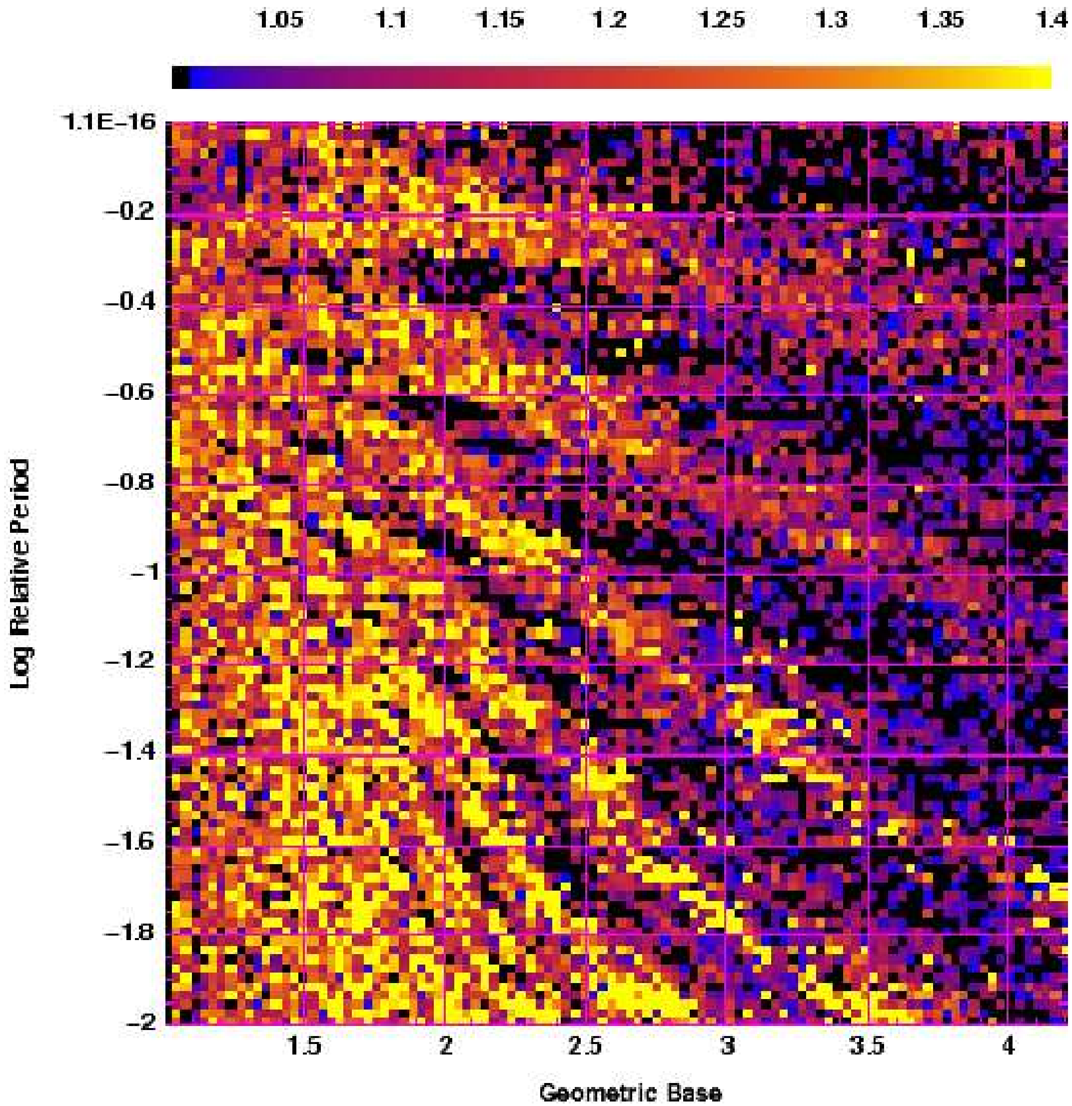

Figure 3.23: *A* for a 12 point ordered geometric lightcurve, as a function of geometric base and period.

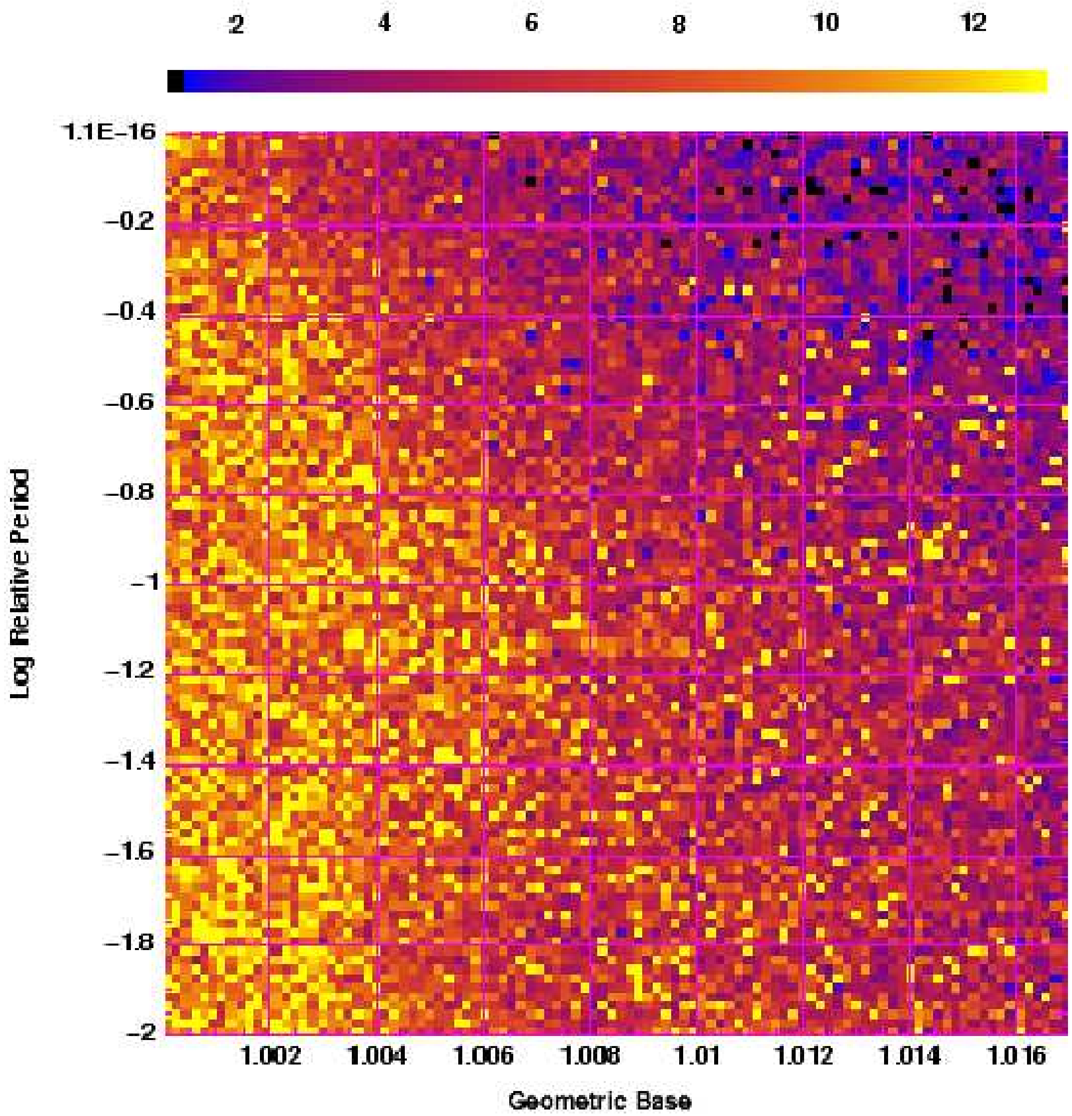

Figure 3.24: A for a 1000 point ordered geometric lightcurve, as a function of geometric base and period.

then that frequency is sampled, but only once. But if this gap is somewhat smaller still, then the highest frequency is sampled several times by the gaps between non-adjacent points. The price is that the very smallest gaps are sampling some frequencies higher than the regime of interest. This means irrelevant frequencies are being acquired at the expense of coverage in the desired range. Hence the position of the optimum base is a trade-off, balancing small numbers of useless high frequencies with the additional useful frequency coverage their observations provide.

This picture also explains the roughly triangular region of optimality in Figure 3.24. For any choice of distribution, including linear or random, the majority of the covered frequencies lie at the mid-range of the period search space. Simply put, there are only a few very short and very long frequencies. This means that mid-range periods have the best frequency coverage, and are the most resilient if the geometric base is increased. It also raises an interesting possibility: since for any distribution the mid-range of frequencies is rather similar, the distinctions between sampling strategies are largely governed by the extreme frequencies. Therefore shuffling a geometric sampling, preserving the individual smallest gaps but changing the set of mid-range gaps, is an interesting idea. It is explored in the next section.

3.7 Randomised geometric placement

Given the favourable properties of a standard geometric sampling, a good question to ask is whether the order of the sampling matters. The choice of spacing may be preserved while modifying the order in which observations take place. This is equivalent to shuffling the gaps between observations. One motivation for doing this is that it allows individual observations to be placed with much more flexibility than an ordered geometric spacing, greatly easing scheduling problems.

Figure 3.25 plots the variation of RMS with geometric base for a 25 point reshuffled geometric spacing. For comparison, the ordered geometric spacing is overplotted (dashes). In general, a randomised geometric spacing has a slightly higher RMS than the equivalent ordered spacing over the range of optimal bases. Figure 3.26 shows the window function for the base with the lowest RMS. It should be noted that each random realisation has its own unique window function, and that for small values of N the RMS values can vary markedly from run to run. However, the range of variation in RMS across the space of possible random realisations shrinks rapidly with increasing N.

The different behaviour between the ordered and randomised geometric spacing is more clearly illustrated in Figure 3.27. The optimal range of bases remains almost the same as for the ordered case. In general, the smoothness (as measured by the RMS value) of the window function of the best randomised geometric spacing is not quite as good as the equivalent ordered spacing. However, the randomised spacing degrades much more smoothly with increasing geometric base — for sub-optimal choices of the geometric base, a randomised spacing out-performs the ordered spacing. This has important implications for observing programmes in which the total number of observations is a highly dynamic quantity which cannot be accurately predicted in advance. In such a context, choosing a randomised geometric spacing would allow the observing programme more flexibility with the total number of observations and their spacings, while seeking to optimise

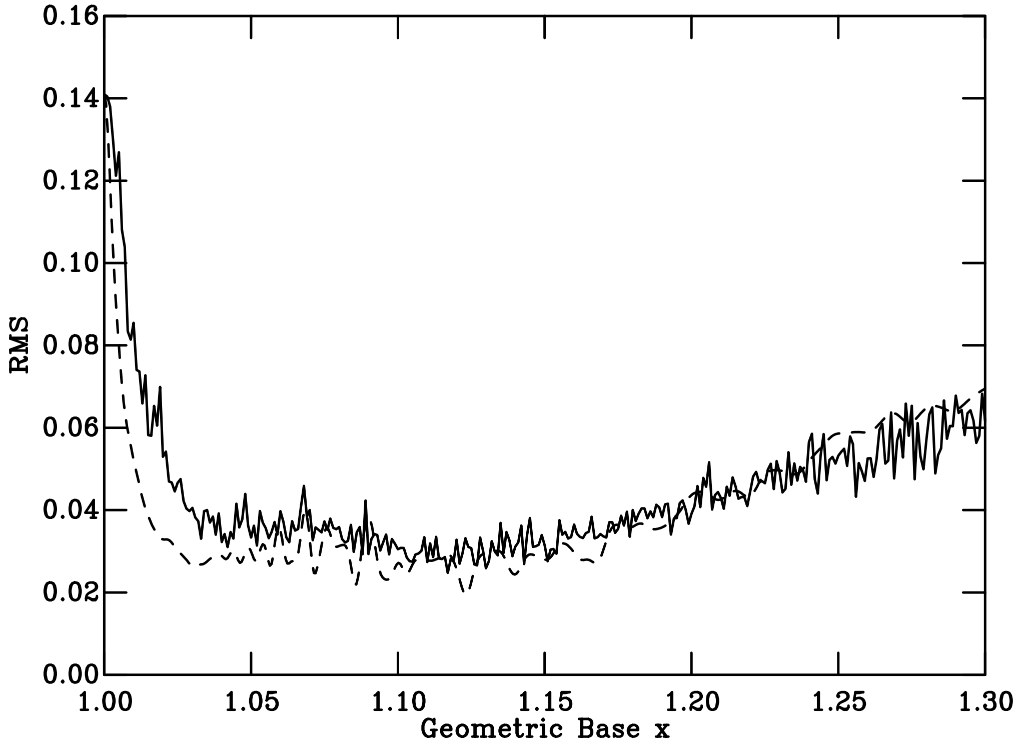

Figure 3.25: RMS as a function of randomised geometric base, for $N = 25$ (solid line). For comparison, the ordered geometric spacing has been overplotted (dashes).

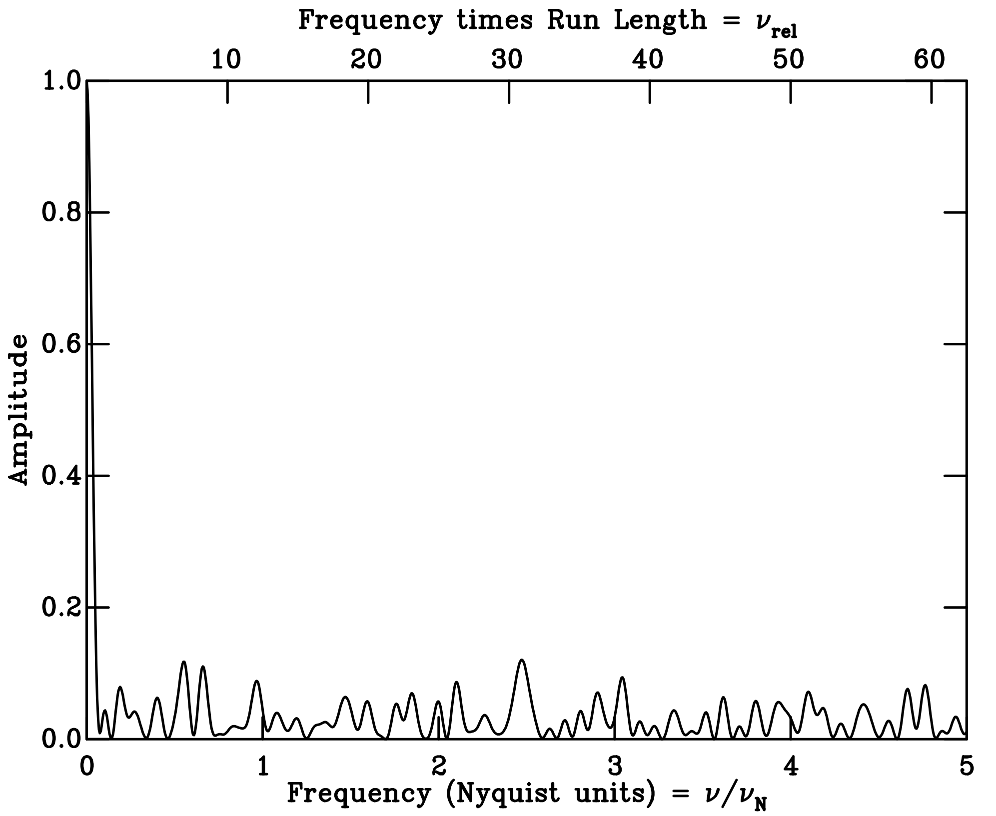

Figure 3.26: A typical window function for an optimum randomised geometric sampling of 25 data points, where $x = 1.128$. The RMS is 0.025.

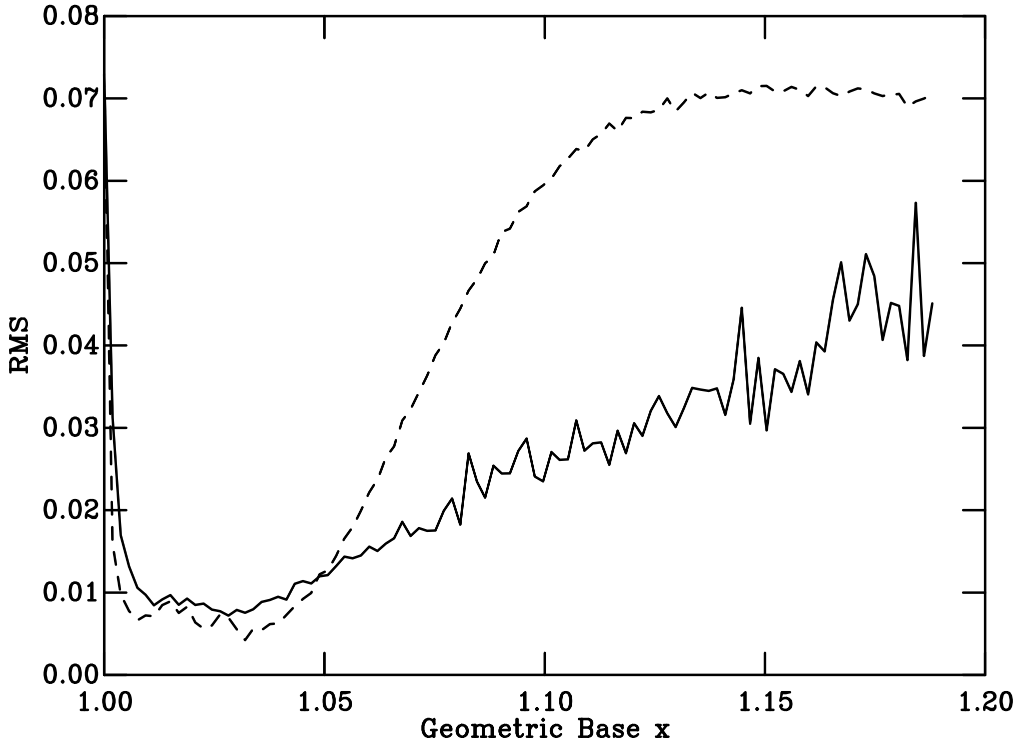

Figure 3.27: Randomised geometric base as a function of RMS, for $N = 100$ (solid line). For comparison, the ordered geometric spacing has been overplotted (dashes).

the schedule based on the current estimate of the run size and duration. Such an algorithm could form the basis of an *adaptive scheduling agent*, an autonomous software entity which encapsulates sampling knowledge and attempts to exploit it in a real world environment. Chapter 4 describes such an algorithm.

3.8 Practical application

Figure 3.28 presents the optimal geometric base, found by minimising the RMS of the window function, for each of a range of randomly respaced geometric sampling scenarios. By taking the period range and run-length of interest, converting them into a relative limiting frequency, and then estimating the total number of observations to be made, the ideal geometric base for different observation regimes can be found. Expressing the limiting frequency in units of ν_{N}, and substituting Equation 3.1 gives

$$\left(\frac{\nu}{\nu_{\mathrm{N}}}\right) = \frac{2T\nu}{N} = \frac{2\nu_{rel}}{N} \tag{3.16}$$

As an example, an astronomer with access to a robotic network is planning a 3 week observing run, searching for periods ranging from 5 hours to 10 days. This corresponds to a minimum relative period of 0.0099, and thus a maximum limiting relative frequency ν_{rel} of 101. If the total number of observations is likely to be around 100, then $(\nu/\nu_{\mathrm{N}}) \approx 2$. Applying the lower curve in Figure 3.28, the ideal sampling base is found to lie at around 1.02.

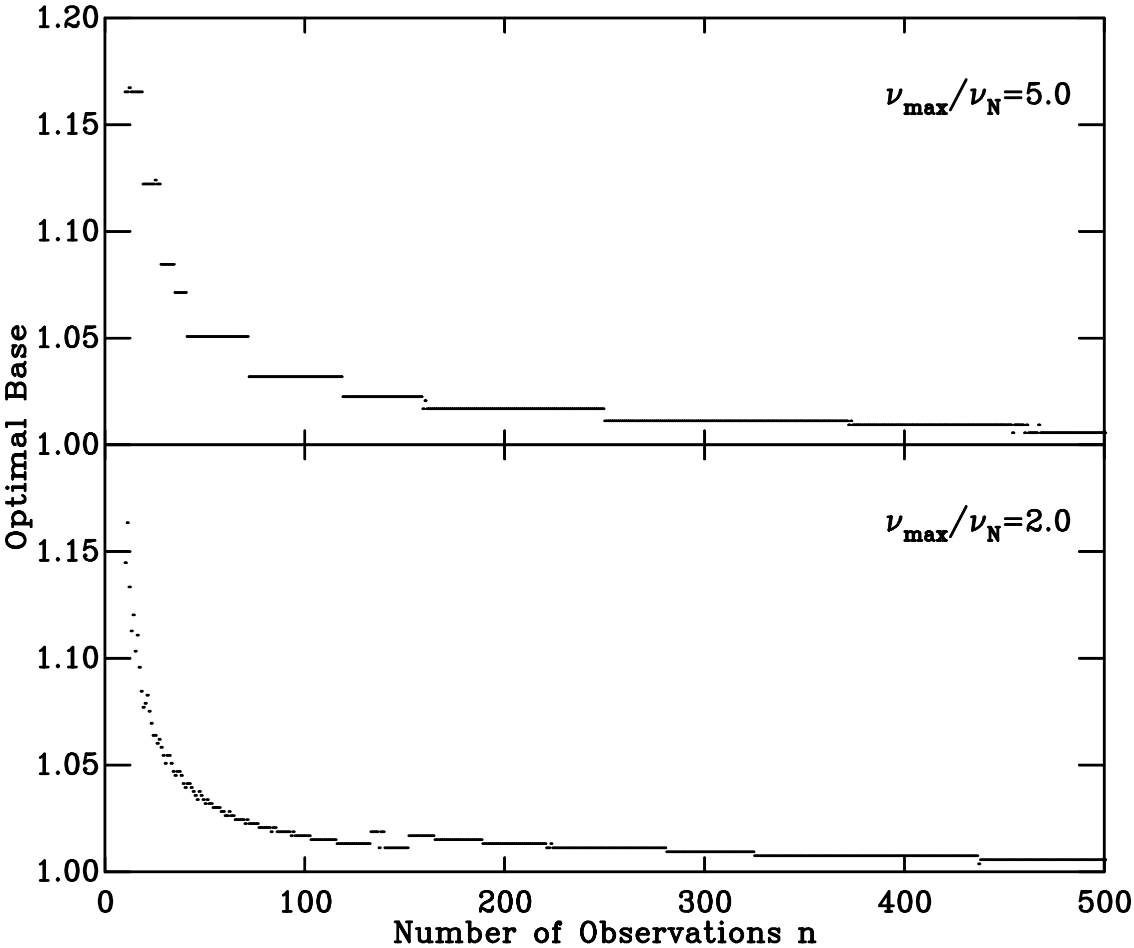

Figure 3.28: Optimal geometric bases, as determined by minimum RMS, for datasets spanning 10 to 500 datapoints. For a relative maximum limiting frequency of 2 (lower panel), optimal bases tend to be closer to linear sampling than an equivalent sampling extending over a much higher dynamic range (upper panel). Simulations out to a relative maximum limiting frequency of 12.5 show almost no deviation from the optimal bases calculated for the $\nu_{max}/\nu_N = 5$ case.

Although Figure 3.28 has, in principle, to be recalculated for each particular value of $(\nu/\nu_{\rm N})$, in practice no change was found in the set of optimum bases for relative maximum limiting frequencies greater than 5, and the optimal minima are sufficiently broad that crude interpolation should suffice for smaller values. If the relative limiting frequency is found to be below 1, then enough datapoints exist to sample evenly. No further optimisation is required.

Thus the observer should apply the following procedure to build an observing schedule. The signal range of interest must be identified, and the likely size of the dataset and its duration estimated. This allows the optimal geometric base to be calculated. Having ascertained the optimal base, Equation 3.10 may be used to calculate a series of observation times. The resulting gaps may then be reordered as desired as the run proceeds, for example to accommodate practical constraints such as weather conditions or telescope maintenance, or to facilitate a single night's intensive coverage during a period of dark time.

3.9 Conclusions

The work in this chapter has sought to answer the question of how best to place a limited number of observations in such a way as to minimise aliasing effects arising from undersampling. A series of metrics for measuring phase coverage (S), cycle count (interconnectivity, N) and the quality of the periodogram (A) have been presented. The properties of various types of samplings have been analysed in these terms, and a detailed understanding of undersampled behaviour developed. By applying a simple geometric sampling rule, the ability to significantly improve the quality of the periodogram has been demonstrated. This placement strategy outperforms previous sampling schemes from the literature. It has the advantage that it is easy to apply at the telescope for datasets of arbitrary size, has calculated empirical solutions for different observing scenarios, and exhibits surprising flexibility arising from the effects of reordering. By careful sampling the improvements that can be made in data quality, or alternatively the savings that may be accrued in telescope time are substantial.

Chapter 4

An Agent for Variable Star Observing

4.1 Introduction

4.1.1 Motivation

In the previous chapter an optimal sampling technique was presented that was based on the geometric spacing of observations. This provides a way to determine the ideal set of observations for a particular period-sampling problem (Saunders et al., 2006b). It was argued that such a series was ideally suited to a multiple-telescope paradigm because of the potential to observe at very specific, widely separated time intervals, and the ability to overcome diurnal breaks in coverage by the judicious use of longitudinally distant telescopes.

However, observations are not guaranteed. In the dispatch model of telescope scheduling users request observations, but it is up to the telescope scheduler to decide whether such requests are honoured. Observations may still fail even when submitted to an entirely cooperative scheduler, due to the possibility of telescope downtime or, more commonly, inclement weather. Thus, any system seeking to reliably implement an optimal geometric sampling technique must in practice deal with the issue of observation uncertainty.

This chapter describes the implementation of an adaptive, dynamically determined algorithm that addresses the practical problem of observing undersampled, periodic, time-varying phenomena. The system is an implementation of a branch of computer science known as *multiagent systems*, a relatively young field related to artificial intelligence. In the multiagent software model, the atomic component of the system is is an autonomous entity capable of making its own decisions called an *agent*. Individual agents interact with one another, communicating with each other in order to achieve a greater goal. The general behaviour of the system is thus the product of the interactions of the multiple agents from which it is composed.

A brief overview of multiagent systems follows, that highlights only those aspects immediately relevant to this chapter. This is followed in Section 4.2 by a discussion of the architecture of the eSTAR multiagent system (Allan et al., 2004a), and the physical network of telescopes called Robonet-1.0 over which it operates.

Section 4.3 describes the design and implementation of the decision-making algorithm of the adaptive scheduling agent. Section 4.4 describes the testing process, including network simulations. The results of the on-sky observing run (managed autonomously by this agent) are presented, and the performance of the agent is discussed in Section 4.5. Finally, conclusions are presented in Section 4.6.

4.1.2 What is an agent?

Perhaps rather surprisingly, the question of how an agent is defined is not as straightforward as it may at first appear. Definitions of agency tend to fall between two extremes. At the most reductionist level, an agent is any automata which behaves as it has been programmed. This extremely broad definition thus includes most computational processes, including ultra-fine-grained processes such as addition. At the other extreme, agents are conscious, cognitive entities. Since no software agents have ever been developed which would fulfill such a criterion, under this definition the entire field of agent research is inadequate.

It would seem that the first definition is too permissive to be usefully applied, while the second is so far beyond the current state of the art that it has little practical use. For the purposes of this discussion, a definition amalgamated from Wooldridge & Jennings (1995) and Huhns & Singh (1998) shall be used here:

> Agents are active, persistent software components, which perceive, reason, act, communicate, and are capable of *autonomous* action in order to meet their design objectives.

4.1.3 Autonomy

One of the properties which makes agents difficult to define is the notion of *autonomy*. The problem arises because there are a number of different kinds of autonomy which may be exhibited by ‘autonomous systems’. The following are described by Huhns & Singh (1998).

- **Absolute autonomy.** An absolutely autonomous agent is one which may choose any actions it likes. In general, absolute autonomy is not a desirable feature of agent systems, because useful agents ordinarily have some purpose envisaged by their designer, which constrains them.

- **Social autonomy.** Coordination with others reduces the autonomy of individual agents. For example, by choosing to queue at a bus-stop, an individual gives up some portion of their autonomy (their freedom to get on the bus first), in order to coordinate with others attempting to board the bus. In this way, the good of the whole is maximised at the expense of the individual. Social autonomy is the case where an agent attempts to coordinate with others where appropriate, but displays autonomy in its choice of commitments to others (e.g. in making the decision to join the queue).

- **Interface autonomy.** To perform useful functions, agent autonomy is typically constrained by an API (application programming interface). Interface autonomy describes the level of autonomy hidden behind the API of an agent — what the agent can choose to do subject to obeying the API. It is therefore autonomy with respect to internal design.

- **Execution autonomy.** The extent to which an agent has control of its own actions is its level of execution autonomy. This flavour of autonomy arises because an agent that is controlled to some extent by a user or other process may appear to be autonomous to other agents, but is clearly less independent than an uncontrolled agent. An example of the constraint of execution autonomy is an e-commerce agent that requests verification from a user before proceeding to complete a transaction (Chavez & Maes, 1996).

- **Design autonomy.** The extent to which an agent design is constrained by outside factors is described by design autonomy. For example, communication with other agents may require an ability to represent beliefs, or for communications to be implemented in a specific language. The design is therefore constrained by these requirements.

4.1.4 Multi-agent systems and other fields

The field of agents overlaps a number of other areas of computer science. It is therefore useful to describe how agents differ from other areas of research.

- **Objects.** Agents and objects seem at first sight to be rather similar. Objects are defined as 'computational entities that encapsulate some state, are able to perform actions, or methods on this state, and communicate by message-passing' (Wooldridge, 2002). However, there are three main differences between the agent and object models. Firstly, while objects may exhibit autonomy over their internal *state* (through the use of public and private variables, for example), they exhibit no control over their own *behaviour* — that is, they do not possess any degree of execution autonomy. Secondly, the object model has nothing to say about reactive, proactive, or social behaviour — all key features of agent systems. Finally, multi-agent systems are by definition multi-threaded, concurrent systems, possessing many threads of control. Critically, an agent encompasses its *own* thread of control. While some form of threading is commonly implemented in object-oriented languages, it is not a requirement of the standard object model.

  This is not to say that agents cannot be implemented in an object-oriented manner. The agent abstraction is simply a higher level view of the system architecture, and is concerned with different issues.

- **Distributed and concurrent systems.** As described above, multi-agent systems are a subset of concurrent systems. Therefore all the problems associated with concurrency, such as deadlock, synchronisation, resource sharing and so on are also issues in multi-agent systems. However, the focus of the two fields is rather different. In particular, multi-agent systems frequently require synchronisation to take place dynamically, at run-time. Also, because entities in concurrent systems typically share a common goal, problems of negotiation and cooperation with other self-interested entities are not a feature of the concurrency field.

- **Artificial Intelligence (AI).** There are two main differences between the agent field and the broader field of artificial intelligence. Firstly, AI is largely concerned with the components of intelligence. These include learning, understanding, planning and knowledge representation. Agent systems are concerned with the integrated final machine — how to build an autonomous, decision-making agent, and interact successfully with other agents to solve problems. It turns out that it is not necessary to solve the (numerous) problems of AI in order to build successful agents. Secondly, multi-agent systems are concerned with problems of *societies* — cooperation, coordination, negotiation, and so on. AI has traditionally taken a much more reductionist approach, concentrating on the individual, and consequently has little to say about these problems.

- **Web and Grid services.** The fields of Web services and Grid computing are concerned with solving the problems of integrating heterogeneous resources and processes in complex, dynamic, massively distributed systems (the Internet and the various Grid environments). They overlap with agent systems in areas such as concurrency, resource brokering, trust, communication, wrapping of legacy components and reaction to changes of environment. However, Web services are much more like objects than agents. They respond to external requests, and do not display behavioural (execution) autonomy. The other difference is situational — agent systems need not be based inside Internet or Grid environments (although this is likely to become the dominant trend within the field). Most of the successful agent projects to date have been bespoke applications situated in very specific environments such as industrial processing and air traffic control, for example (Jennings et al., 1996; Ljunberg & Lucas, 1992).

4.1.5 Reactive architectures

Reactive or situated approaches to agent-building arise from a rejection of the classical AI approach of logical or symbolic reasoning. Authors such as Brooks (1991) and Rosenschein & Kaelbling (1996) propose that the *environment* in which an agent system is situated is a critical feature of intelligent, rational behaviour, and cannot be divorced from the internal agent architecture. In particular, intelligent behaviour can be viewed as a function of the interaction of an agent with its environment. This radical idea has given rise to the field of *emergent systems* — systems built from individual components which do not exhibit reasoning capabilities, but which interact with each other and the environment to produce complex, intelligent behaviour. This idea of 'intelligence without reason' (Brooks, 1991) takes as its inspiration examples from the natural world such as ant colonies and bee societies, as well as sociological ideas such as theories of group behaviour (such as crowd motion) and large scale trends across society (such as the emergence of fashions or fads).

In physics, it has long been understood that non-trivial, emergent behaviour can arise from systems governed by relatively simple rules. Macroscopic phenomena such as phase changes arise as a consequence of the dimensionality of the system and the form of the interactions between components, irrespective of the individual properties of those components. Thus we observe phase changes in liquids and gases, but also in magnetism and the theory of inflationary cosmology (Guth, 1981). More recently, statistical physicists and economists have begun to apply the study of large-scale systems to theories of society, and have discovered phase changes in properties such as marriage statistics (Ormerod & Campbell, 1998) and so-called ‘tipping points’ - the way that abrupt social changes take place, in an epidemic-like manner (Schelling, 1978; Gladwell, 1998).

4.1.6 Pros and cons of reactive agents

Adopting a reactive agent architecture has a number of attractive features. The architecture is conceptually simple and computationally extremely economic (because there is no computational reasoning taking place). It is a robust architecture, because multi-agent systems constructed using this model tend to be tolerant in the face of dynamic, uncertain environments (the loss of a single ant has a negligible impact on the performance of the colony as a whole). It neatly sidesteps the issues of individual agent intelligence by considering instead arising emergent behaviour as the mechanism for intelligent action. Finally, it is an elegant and novel architecture which draws parallels with a wide variety of successful distributed systems in nature.

However, there are some serious limitations to the reactive approach to agency (Wooldridge, 2002). Perhaps the most serious is that, in general, reactive agents do not model their environments, and consequently cannot make decisions based on history or predict the future. Consequently all actions are inherently short-term, and are based entirely on local information (information about the current state of the environment). Additionally, purely reactive agents do not learn from experiences, and therefore cannot improve performance over time. The other major problem arises from the reliance on emergent properties to solve the problems of the multi-agent system. This overall emergent behaviour is difficult to predict, and consequently engineering agent systems to fulfill specific tasks is hard, potentially reducing to a process of trial and error that can be uncertain and time-consuming. Indeed some researchers have claimed that emergent properties cannot be derived, even from a perfect understanding of the situation, with less computation than direct simulation (Darley, 1994).

4.2 The eSTAR project

The eSTAR project[1] (Allan et al., 2004a) is a multi-agent system funded as part of the UK e-science programme, and aims “to establish an intelligent robotic telescope network” (Allan et al., 2006). The project is a collaboration between the Astrophysics Research Institute of Liverpool John Moores University, the Astrophysics Group at the University of Exeter, and the Joint Astronomy Centre in Hawaii.

There are two main types of agent in the eSTAR network. *User agents* act on behalf of an astronomer, and are responsible for carrying out individual science programmes. *Node agents*, sometimes also called *embedded agents*, provide the interface to observational resources such as telescopes or astronomical databases. A user agent typically seeks to fulfill its scientific mission by interacting with multiple node agents and negotiating for suitable observations. Figure 4.1 illustrates the high-level view of the system.

[1]e-Science Telescopes for Astronomical Research

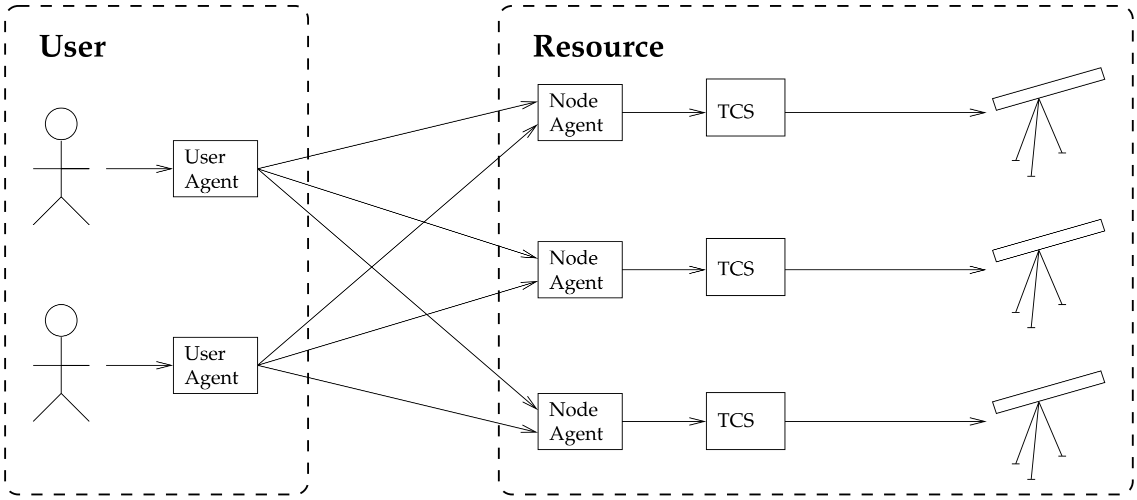

Figure 4.1: A high-level view of the eSTAR architecture. User agents run an observing programme with parameters specified by a scientist. Node agents provide the interface to the telescope control system (TCS) which implements observation scheduling as well as lower-level telescope control functions.

4.2.1 Communication architecture

eSTAR is a peer-to-peer software network that overlays an underlying physical collection of disparate telescopes. The architecture is based on the software as a service model, where individual code components are made accessible as web services. Much of the communication that would normally be handled internally by variable passing is explicitly passed using transport protocols such as SOAP or lower-level socket connections. This approach provides a generalised abstraction to the physical location where the code runs, allowing the code components to be arbitrarily distributed across a physical network. The disadvantage of the approach is that interprocess communication is more complex than in a traditional monolithic code, and the possibility of communication failure makes it impossible to ever guarantee successful message-passing, necessitating additional layers of error-handling code.

In the sense that logical components may be physically separated the architecture is highly decentralised. However, at the level of individual science programmes the system is centrally controlled and managed, with a single instance of a user agent managing a single science programme. Individual user agents interact with many node agents, often by global broadcast (one-to-many), but no currently implemented user agent communicates with any other. Similarly, individual node agents are unaware of one another, but typically interact with many user agents on a one-to-one basis[2]. Thus in general one-to-many (user agent—node agent) and bi-directional one-to-one communication between members of these two distinct agent populations is typical (see Fig. 4.2). This type of communication scales well because each new agent adds only *n* new communication paths, where *n* is the number of agents in the *other* population. Thus the communication load scales approximately linearly with additional agents (contrast this with the polynomial increase in

[2]There is no architectural limitation that would prevent node agents from communicating. A proof of concept experiment, where the FTN and UKIRT node agents communicated directly, has in fact been performed (Allan 2007, priv. comm.).

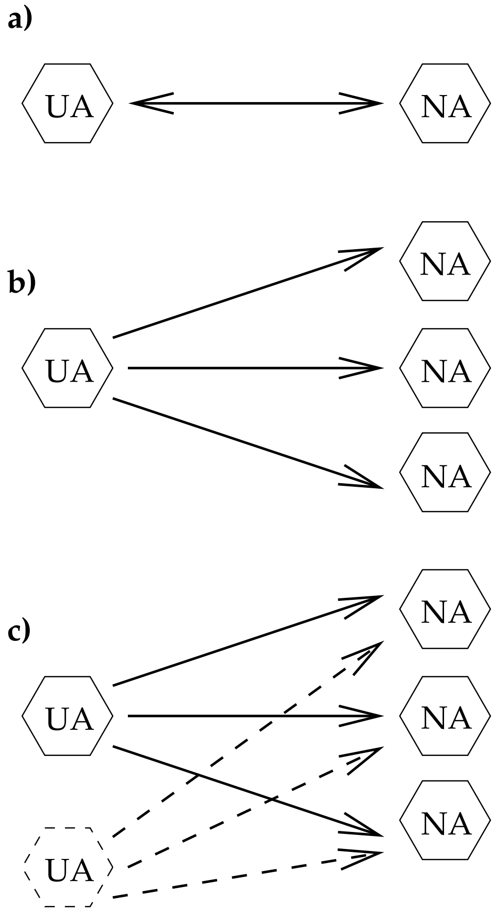

Figure 4.2: Types of communication in the eSTAR system. User agents and node agents can both send point-to-point messages (*a*). User agents can also broadcast the same message to all node agents (*b*). Adding new agents to the pool increases the number of communication paths by *n*, where *n* is the number of agents of the *other* population (*c*).

communication overhead in a general many-to-many agent system!).

4.2.2 User agents and node agents

User agents in eSTAR are rational and self-interested, and seek only to fulfill their science goals. Internally they are mostly implemented as partial-plan agents, with a set of simple behaviours specified for different situations. Such behaviours include negotiating with and selecting node agents for observations, requesting follow-up observations, sending emails or SMS text messages to an astronomer, and responding to real-time alerts. The adaptive scheduling agent (ASA) is an exception. It implements a combination of hardwired behaviour for particular scenarios, and a dynamically calculated observing strategy based on successful observations to date.

A node agent does not itself implement self-interested behaviour, but it does convey the state of the telescope scheduler that it represents. The node agent provides an indication of the scheduler's inclination to perform observations through score responses to individual user agent queries. The node agent also conveys the ultimate outcome of observation requests. In this way the node agent appears to a user agent to be both rational and self-interested. It provides information about its 'state of mind' and indirect information about its true 'attitude' by the actual observing actions it performs. The point is that although user agents currently perform no modelling of node agents in terms such as beliefs, desires and intentions, such a modelling is entirely feasible.

4.2.3 Robonet-1.0

The three Robonet-1.0 telescopes (Bode et al., 2004) are 2 m optical telescopes of identical design manufactured by Telescope Technologies Limited. The Liverpool Telescope was the prototype instrument, and is situated on La Palma in the Canary Islands. The Faulkes Telescope North is located on the island of Maui, Hawaii. The Faulkes Telescope South is located at Siding Springs Observatory in Australia. The telescopes run identical dispatch schedulers, and accept observation requests from users as well as eSTAR user agents. The Faulkes telescopes also dedicate part of their available time for use by schoolchildren around the world. Observing time for astronomers on these telescopes is allocated by a time-allocation committee based on scientific merit.

The Robonet-1.0 telescopes are completely autonomous. They cannot be compelled to take the observations of a particular user agent (or user). The telescope scheduler seeks to optimise the usage of the available time for all observing projects, and as such reserves the right to refuse observing requests, or to default on existing queued requests. This situation is very similar to the paradigm example of a computational resource in Grid computing, which is typically owned by a service provider anxious to retain control of their hardware. Nodes in a computational grid retain the right to refuse job requests, and to default on existing jobs subject to quality-of-service constraints. The constraint of resource autonomy is vital, because it allows service providers to join the network safely — there is nothing to lose. This is a fundamental tenet of the eSTAR approach, and for this reason the network is sometimes described as an 'Observational Grid' (Allan et al., 2003).

4.2.4 Implementation of the eSTAR network

The eSTAR code is written entirely in object-oriented Perl (Wall et al., 2000). The eSTAR agents communicate with one another using the XML dialect RTML (Pennypacker et al., 2002; Hessman, 2006b), with network alert messages implemented using the VOEvent Protocol and Transport standards (Seaman et al., 2005, 2006). RTML provides a standardised and carefully designed set of elements that aims to allow an observation request to be fully specified. The actual transport protocol used for conveying the XML from node to node is SOAP (Simple Object Access Protocol) over HTTP. SOAP is a well-known application layer protocol designed for the transfer of data between web services. It provides a robust way to send messages reliably between network nodes.

The HTN protocol

If RTML is the language of the eSTAR network, then the HTN protocol (Allan et al., 2006) is its social customs. It specifies the steps that any HTN-compliant resource should pass through in order to negotiate the use of that resource with a user agent. Figure 4.3 illustrates the process. The first two steps of the protocol are optional, and are not implemented by eSTAR. The first step

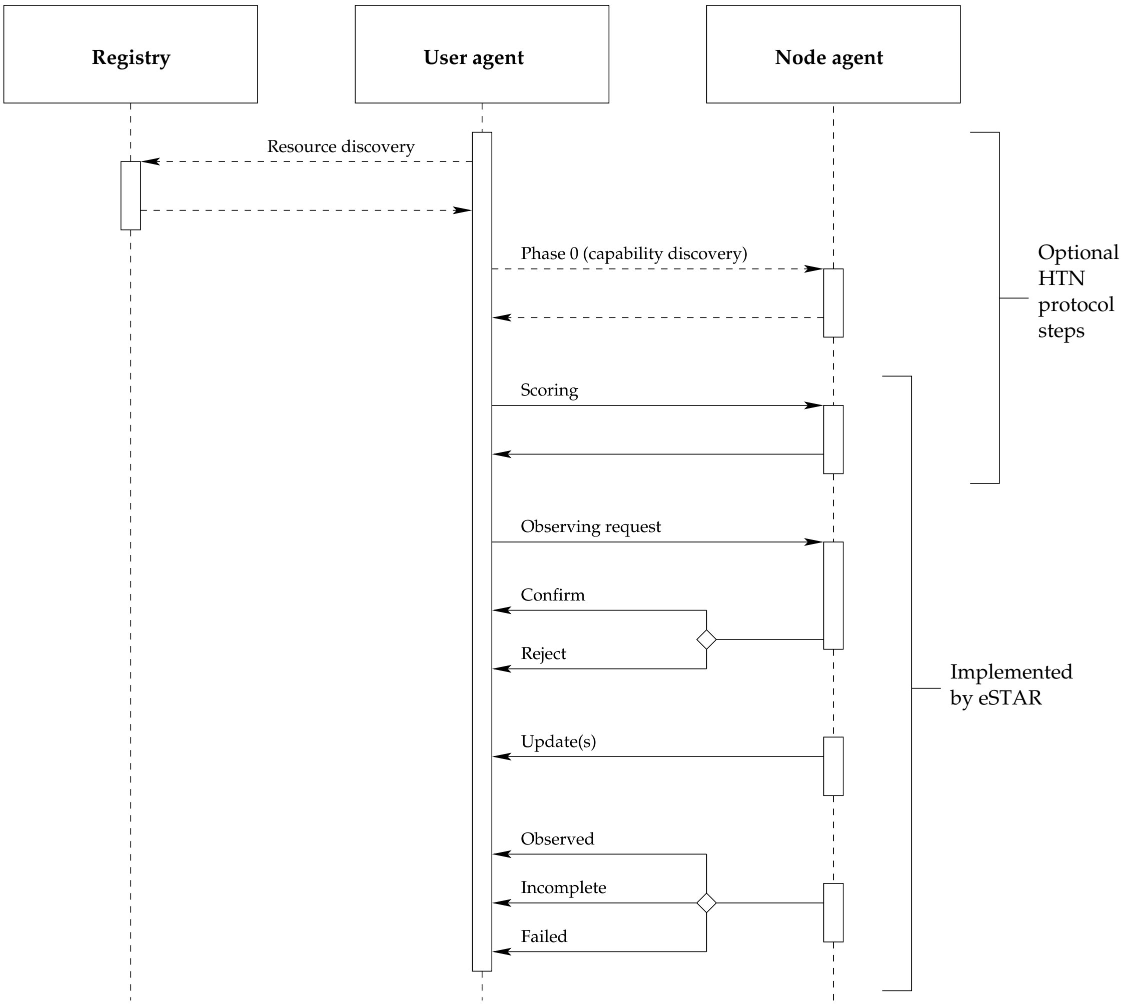

Figure 4.3: Sequence diagram illustrating the eSTAR implementation of the HTN protocol, for a single agent-agent negotiation. Communication begins with a score request from the user agent to the node agent (dashed arrows indicate optional protocol steps not presently implemented in the eSTAR system). A non-zero score response leads to an observation request, which is formally acknowledged. The remaining steps are performed asynchronously. Update messages are sent for each frame of the observation request (an eSTAR convention — the content and triggering of updates is not defined by the protocol), followed by a concluding message when the request is completed. The message type indicates whether all (‘observed’), some (‘incomplete’) or none (‘failed’) of the data were successfully acquired.

allows for the dynamic discovery of new resources by querying an external registry service of some kind. The second step ('phase 0' in the terminology of the observatories) allows resources to return capability information that could prove useful to an agent seeking a particular resource configuration, for example the presence of a spectrograph or the ability to observe in the infrared. The eSTAR sequence begins with the third step, an optional 'score request', which is an RTML message from a user agent that specifies a possible future observation. The telescope responds with a 'score reply' document, which provides a score between 0 and 1 indicating some sort of likelihood-of-success function for the observation. A zero is a special value that indicates the telescope is certain the observation will not succeed. This is typically because the observation parameters are not resolvable (e.g. the object is not visible at the requested time), or because the telescope is aware of additional considerations such as engineering downtime that will render the observation impossible.

Based on the score, the user agent may decide to submit the observation request to the telescope. An acknowledgement (or rejection) message confirming the agreed observation parameters is sent back to the agent.

The remaining messages are sent asynchronously. 'Update' messages are not defined or mandated by the HTN standard. Because a single observation request can involve multiple exposures, potentially widely-spaced, the eSTAR implementation chooses to send an update message back to the user agent after each successful frame. This allows the user agent to gain timely feedback about the progress of individual components of the observation.

Finally, the observation request resolves. The HTN protocol guarantees observation termination, either because all frames were completed successfully, because some of the frames completed successfully, or because at the expiry time of the request none of the frames were acquired. These situations are indicated by the 'observation', 'incomplete' and 'fail' documents, respectively.

Having described the eSTAR network topology in terms of software, hardware, and inter-agent communication, I now move on to discuss the adaptive scheduling agent itself.

4.3 Building an adaptive scheduling agent

This section describes the details of the eSTAR adaptive scheduling agent (ASA) that implements the theory of optimal sampling discussed in Chapter 3. Section 4.3.1 discusses the internal architecture of the ASA, and provides the details of how messages are generated and received. This lays the groundwork for the explanation of the adaptive decision-making capabilities that give the agent its 'intelligence', investigating the assumptions and design principles (4.3.2) and then the design and implementation of the core algorithm itself (Sections 4.3.3–4.3.7).

4.3.1 Architecture

Figure 4.4 shows the detailed architecture of the adaptive scheduling agent. The core decision-making algorithm is a multi-threaded Perl process that determines when to observe, translates between abstract theoretical timestamps and real dates and times, and monitors the progress of the

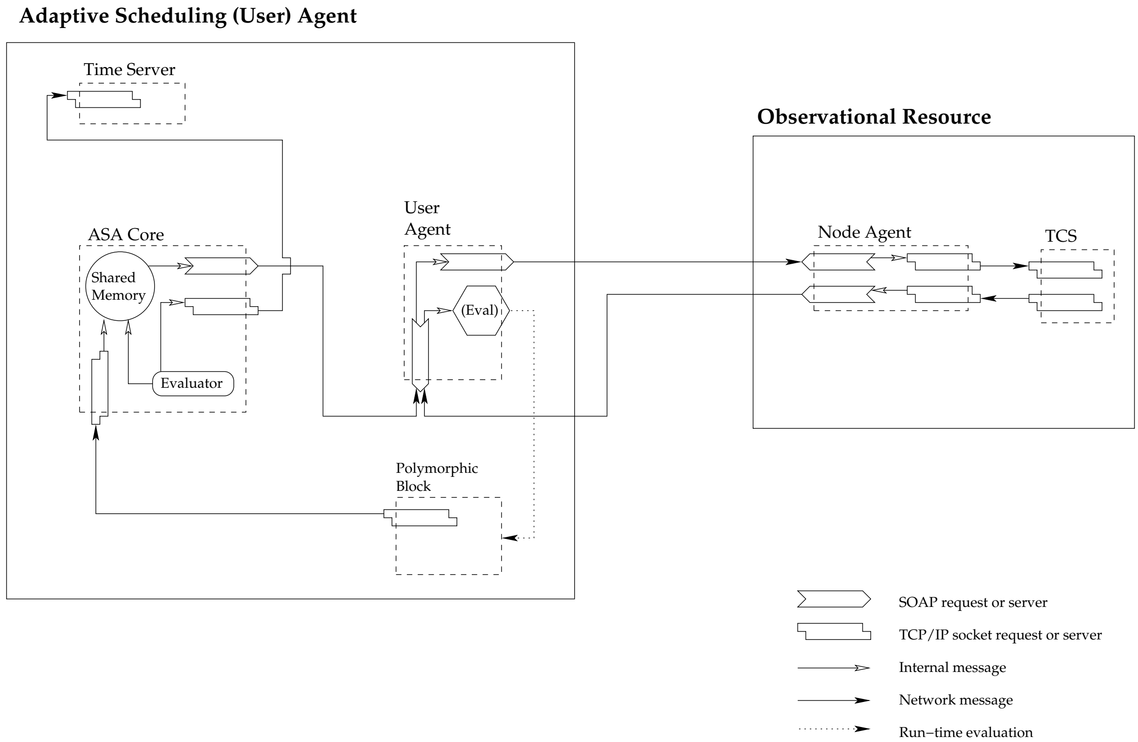

Figure 4.4: Architecture of the adaptive scheduling agent. The multi-threaded ASA core implements the adaptive evaluation engine, a SOAP submission mechanism for making observation requests, and a TCP/IP socket listener for receiving observation feedback information. Date/time information is abstracted by a time server, which can be accelerated to allow fast simulations to be performed. Actual RTML document construction and HTN protocol negotiations are handled by the user agent, a web service that implements a SOAP server for submitting observations and receiving asynchronous messages from the node agent. Asynchronous messages trigger a run-time code evaluation based on the observation type, allowing arbitrary response behaviour to be defined for different observing programmes. In this case, the polymorphic block strips pertinent information (primarily the observation timestamp) from the incoming message and pipes it to the dedicated socket listener in the ASA core. In this way the agent receives observation feedback as the run progresses, allowing direct optimisation of the future observing strategy.

observing run. Internally, the decision-making process uses abstract timestamps, normalised such that the run begins with the first successful observation at timestamp 0, while timestamp 1 represents the requested run end-point (e.g. 5 days in the actual performed experiment). Abstracting the times in this way makes the decision-making process of the agent much easier to understand, because the log files, which contain the detailed record of the agent's train of thought, can be directly compared with the gap spacings predicted by theory.

The agent's current knowledge of the run is stored in a multi-level hash that is shared between all threads. Updates to this data structure by individual threads must be carefully co-ordinated to avoid race conditions. The ASA core runs three threads of control. The *evaluator thread* polls the memory at regular intervals. It tracks the statistics of the run (how many observations have succeeded, for example), and takes action (in the form of new observation requests) depending on circumstances. Observations are requested by pushing the parameters onto a synchronised data structure called a *thread-safe queue*[3]. The *submission thread* continuously monitors the queue, and dequeues the parameters to make an observation request. A third thread, the *listener thread*, monitors an internal socket connection for information about completed requests, and updates the shared memory as necessary.

Observation requests are made by remote method invocation of the user agent web service running on the same machine. As previously described, the user agent makes scoring requests of each telescope (by way of the node agent embedded at each site), and then picks the highest non-zero score. A 'request' document is then submitted. Under normal conditions, a 'confirmation' document is returned. This indicates that the observation request has been accepted by the telescope and is now queued.

Eventually the observation request is resolved. Although multiple exposure observation requests are possible, the ASA always requests single observations, as this maximises the amount of scheduling control retained by the agent. Either the observation was successful, and an 'observation' document is returned, or the observation did not succeed, and a message of type 'fail' is returned. The return document is received by the user agent web service, and scanned for observation type. The agent compares the type to the set of types it knows how to deal with, and if there is a match, the code for that type is dynamically loaded and executed. This plug-in architecture allows custom actions to occur for different types of observation, allowing a single instance of the user agent web service to handle RTML marshaling for many distinct observing programmes.

The algorithmic block for the ASA extracts the critical pieces of information from the incoming RTML document. These include the requested start time, and if the observation was successful, the actual start time of the observation, gleaned from the FITS header included in the 'observation' document. Any error messages are also picked up here. This subset of key information is then transmitted by a simple socket connection to the listener thread of the ASA core code. In this way the loop is closed.

[3] `http://perldoc.perl.org/Thread/Queue.html`

4.3.2 Principles of algorithm design

The environment in which the adaptive scheduling agent operates is inherently uncertain. This fact was the fundamental design constraint from which the desirable attributes of the agent were derived. Four principal ideas drove the design.

- **Robustness.** Building software that will operate reliably over a network is complicated. Therefore a guiding principle was to keep the algorithm as simple as possible, to minimise the possibility of unforeseen interactions between system components. With a distributed system it must be assumed that any remote part of the system can fail at any time, either temporarily or permanently. The algorithm needs to be able to make a reasonable attempt to continue under such circumstances, and to 'do the right thing'.

- **Cynicism.** Although all the agents in the eSTAR system are under our control, and so in principle trustworthy, in practice it is much safer to assume nothing binding about events or entities external to the main code. The benevolence assumption, namely that agents may act for the greater good of the system, at their own expense, is not valid here. This is because the telescope schedulers, and by proxy the actions of the node agents, are not under the control of eSTAR. Their goal is to optimise the schedule in some internally satisfactory way, which may or may not coincide with the goals of the user agent. This is most important with respect to the scoring information returned by the observing nodes. Because there is no penalty to the node agent for providing an inaccurate score, there is no compelling reason for the user agent to trust that value. Even if we believed the node was acting in good faith, we still have no *a priori* idea of the accuracy of the information being supplied. Therefore external information should be considered, but not relied upon, and in general we need to consider whether we can adequately handle the worst case scenario (being supplied with false information).

- **Stability.** A poor algorithm would require chains of specific events to be successful in order to achieve its design goal. If the future is highly uncertain, then an algorithm that is reliant on that future to succeed is risky and fragile. Ideally, the performance of the algorithm should degrade gracefully as the environment becomes more hostile, so that it continues to make a best effort to succeed. In practical terms, this means that partial runs need to be optimal, i.e. that whatever the agent has achieved to date needs to be the best set of observations that it could make under the circumstances, since we cannot guarantee future observations at the temporal positions we specify. This is not straightforward. The optimal sampling technique described in Chapter 3 says nothing about how to go about the practical business of acquiring observations. It implicitly assumes all observations will be successful.

- **Adaptability.** Finally, the agent needs to constantly evaluate the current state of the run, and the observing conditions, and alter its behaviour as necessary. A simple example illustrates this point. The agent is aiming to acquire a set of observations that are spaced by both small and large gaps, of particular sizes. If the agent has acquired many small gaps, how should it alter its behaviour to try to achieve larger gaps? What should it do in the opposite

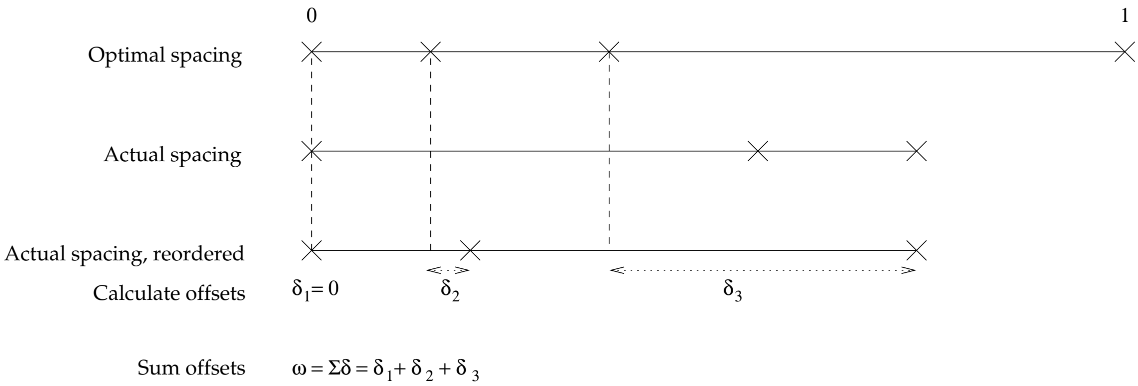

Figure 4.5: Calculating the optimality, *w*, of a set of observations. The optimal and actual sets of gaps are sorted, smallest to largest, and each gap compared. The sum of offsets is *w*.

case? Indeed, how should it adapt to the general problem of a set of gaps that are not quite correct? Another plausible scenario is that of a telescope that consistently provides low scores, but still manages to complete observations successfully. How should that be handled? The agent programming approach explicitly assumes uncertainty and change, and defines a feedback cycle to interact with the environment. How to successfully implement useful adaptive behaviour is therefore of critical importance for the success of this approach.

4.3.3 Defining optimality

If an agent could place observations at any time and be guaranteed of success, then the choice of observations is clear: they should be placed with the gap spacings indicated by the optimal sampling. However in the normal operating environment many observations can fail. When an observation eventually does succeed, the gap between that observation and the last successful observation is unlikely to be of an ideal length — but it could be close.

What the agent requires is some unambiguous way to determine how well its completed spacings compare to the optimal set of gaps. It is not possible to simply compare the two sets of gaps and 'tick off' perfect gaps as they are obtained, because even a 'correct' observation is not precisely located. Telescope overheads mean that in practice an acceptable window for the observation must be provided, and the observation can take place anywhere within that window. Some sort of fuzzy criterion could be used, but this must be explicitly defined and is somewhat arbitrary.

The *optimality criterion*, *w*, is defined by the following simple series of operations[4].

1. Order the set of optimal gaps, from smallest to largest.
2. Order the set of obtained gaps, from smallest to largest.
3. For each obtained gap, find the offset of that gap from the optimal gap at that position in the set.

[4]Note that this is just one of many possible choices of optimality function. This particular formulation was chosen because of the desirable properties outlined in the text.

4. The overall optimality, w is the sum of these offsets.

Figure 4.5 illustrates this process. The optimal gap set is expressed in theoretical time, i.e. so that the run begins at timestamp 0, and the final observation occurs at timestamp 1. The obtained gaps are scaled to the same units. If the set of actual timestamps is in perfect agreement with the optimal set, then the value of the optimality metric is 0. There is no upper limit on the value since the length of the actual run can exceed that of the optimal sequence.

Note that the simplicity of this approach is only possible because of the reordering property of the optimal sequence. This allows the gaps to be compared in order of size, regardless of the actual observed ordering.

4.3.4 Choosing the next observation

The optimality function allows potential future observation timestamps to be compared. The agent seeks to minimise the value of the optimality, which will increase monotonically as more observations accrue. The question is, given the existing set of observations, what new timestamp will achieve this? Since time is continuous, in principle there are an infinite number of possible futures to choose between. Once again it is the ability to reorder optimal observations that allows some elegant simplifications to be made. Firstly, if all reorderings of a given optimal series are considered equally optimal, then the set is degenerate with respect to reordering: for any set of gaps we need only consider one possible reordering. The most straightforward is to place the gaps in order of size. Secondly, it is apparent that any new observation should be placed in such a way as to exactly achieve one of the optimal gaps, since anything else would immediately increase the value of w for no gain. This insight drastically reduces the search space of possible timestamps.

The situation can be illustrated as follows. If there are ten observations in the optimal series, then there are nine optimal gaps that the algorithm is aiming to achieve. If it has already achieved four of them, then there are only five possible future timestamps that need to be considered in order to gain another optimal gap. Formally, for n observations, the ith gap of a possible $n-1$ can be created in no more than $n-1-j$ ways, where j is the number of optimal gaps already obtained.

A third consideration cuts the problem down still further. Consider the situation at the very start of the run. The first observation is requested immediately. Once the agent has a first observation, it needs to consider where to go next. The agent has a choice of $n-1$ ideal timestamps, each of which corresponds to one of the $n-1$ optimal gaps. No gap is *intrinsically* better than any other. In the best case, the gap will be exactly optimal, and the value of w will remain 0. For any other outcome, w will be increased. However, the crucial point is that optimality is calculated by comparing gaps. This means that the magnitude of the increase in w accrued from any new observation is directly proportional to the distance of that gap from its comparison gap. Because we choose to order the comparison gaps from shortest to longest, this means that large gaps early on are penalised by relatively large increases in w. In fact, at this early stage in the run, even a large gap corresponding to one of the required optimal gaps will increase the value of w, because it will be compared to a small gap. Thus the agent is driven to choose the *shortest* optimal gap.

This resultant behaviour is an important feature of the optimality definition. It makes good practical sense to go for the shortest useful gap at any given time, because the agent must wait until that timestamp before it can discover whether the observation actually happened (and hence the required gap collected). Going for a longer gap is much more risky, because the agent is aiming to make the largest number of well-placed observations possible in the finite space of the ideal length of the run. This means it is imperative that the loss of run-time to the uncertainties of the observing process is minimised.

However, it is not sufficient to always simply track the shortest 'uncollected' gap, observation after observation. This would indeed be the best strategy if observations were always guaranteed to succeed, but in practice two types of forced deviations from optimality can be expected to occur in a real observing environment. The first is that some timestamps are known prior to observation to be inaccessible, and hence are guaranteed to fail (e.g. because it is day, the target is not visible, the target is too close to the moon etc.). This information is passed back to the agent in the form of a score of 0. The second is that some fraction of observations that were expected to succeed will in fact fail 'on the night', due to uncertainties in the observing process itself (weather, telescope load, etc.).

These failures force the creation of sub-optimal gaps in the observing pattern. The specific shape of this observed pattern drives the choice of next gap. This is because sub-optimality is defined as a continuous spectrum — a gap of a similar size to an optimal gap is 'more optimal' than a gap of a very different size. The effect of failed observations is to increase the average size of gaps in the observed set. Although these larger gaps were not planned, they can nevertheless be useful, since they will be nearer in size to the larger gaps in the optimal set. It may therefore make sense for the agent to aim for a smaller gap, because this will shunt the larger gaps along in the optimality comparison, improving the overall optimality of the series. In other situations however, it may make sense for the agent to choose a slightly longer gap, even if it has not achieved all the shorter gaps preceding it, because too many short gaps will worsen the optimality of the longer end of the run.

A simple worked example demonstrates the functioning of the algorithm. Figure 4.6 describes an idealised observing period, divided into integer timesteps between 0 and 20 for simplicity. Assume the agent seeks a six point, approximately geometric sequence (Step 1)). The set of gaps G_{opt} that describes this sequence has a perfect optimality of 0, by definition.

Imagine that some observations have already been obtained. Step 2 shows the position of three prior observations. One of them is in a sub-optimal location, giving rise to a non-zero optimality. The agent's problem is to decide where to observe next. It considers each possible optimal gap in turn from the set G_{opt}. In steps 3.1 and 3.2 gaps of size 1 and 3, respectively are considered. The evaluation of w for these steps requires gap reordering from smallest to largest, and the equivalent reordered timestamps are shown in the figure. The remaining gap choices do not require reordering, since they are larger than any existing gap.

Looking at the value of w evaluated for each potential future gap, it is found that of the possible choices, a gap of 3 (evaluated in step 3.2) is the most optimal. This gap is therefore selected, and the observation placed (step 4).

So of the five choices, the lowest (best) optimality comes from choosing the second-shortest available gap, which in this case is the shortest uncollected gap. This is not an explicitly prescribed behaviour; rather it is a natural consequence of minimising structure across the whole run, and arises as a direct consequence of the optimality definition.

4.3.5 Algorithm summary

The continuous optimality function provides the agent with the adaptability it needs to respond to the changing landscape of the run-in-progress. The function automatically penalises the agent for choosing too many short gaps, but is similarly hostile to the chasing of large gaps when there are shorter, more valuable alternatives. In this way a balance point is determined that unambiguously defines the best option for the agent at any point in the run. The deceptively simple rule sequence for finding the optimality implicitly utilises the reordering property of optimal sequences, maximises the agent's cautiousness in the face of uncertainty, and provides a computationally cheap and scalable way to unambiguously calculate the best choice of action at any point. It is also stable — a snapshot of the run at any point in time is optimal for that subset of points, and minimises the effects of observation failure on the sequence. In the case of perfect observing behaviour (no failures), the optimal sequence is naturally recovered. Importantly, the degree of failure exhibited by the network can also change dynamically without adversely affecting the algorithm, because it makes no assumptions about the stability of the network, and makes no attempt to model the behaviour of the telescopes on which it relies.

4.3.6 Submission behaviour

Observation requests are made with a start and end time set to 15 minutes before and after the actual required time. This gives the telescope a 30 minute window of opportunity in which to place the observation. The rationale for these timing constraints was that a 15 minute deviation from the required start time was quite acceptable, while an extremely specific observing request would be less likely to be successful. If the agent receives a score of 0 from every telescope for a single observation request, then it pauses further submissions for a period of 30 minutes (the exact length of time is arbitrary). Although theoretically the agent incurs no disadvantage from continuously requesting the observation, in practice the load placed on the network infrastructure and telescope server is not justified, and the agent has little to gain from such behaviour. The agent will continue to submit requests at 30 minute intervals until a request receives a positive response. Once a request is queued, the agent moves into a waiting state, and will not submit further observations until it receives a return message indicating the success or otherwise of the observation. When the return message is received, the next observation timestamp is determined, and the cycle repeats.

4.3.7 Ideas not used by the algorithm

It is worth emphasising a number of things that the agent does *not* do. The data are not examined, so the A metric (described in Section 3.2.3) is never used. This is by design; observing the Fourier transform of the data during the experiment and changing the sampling behaviour based on the

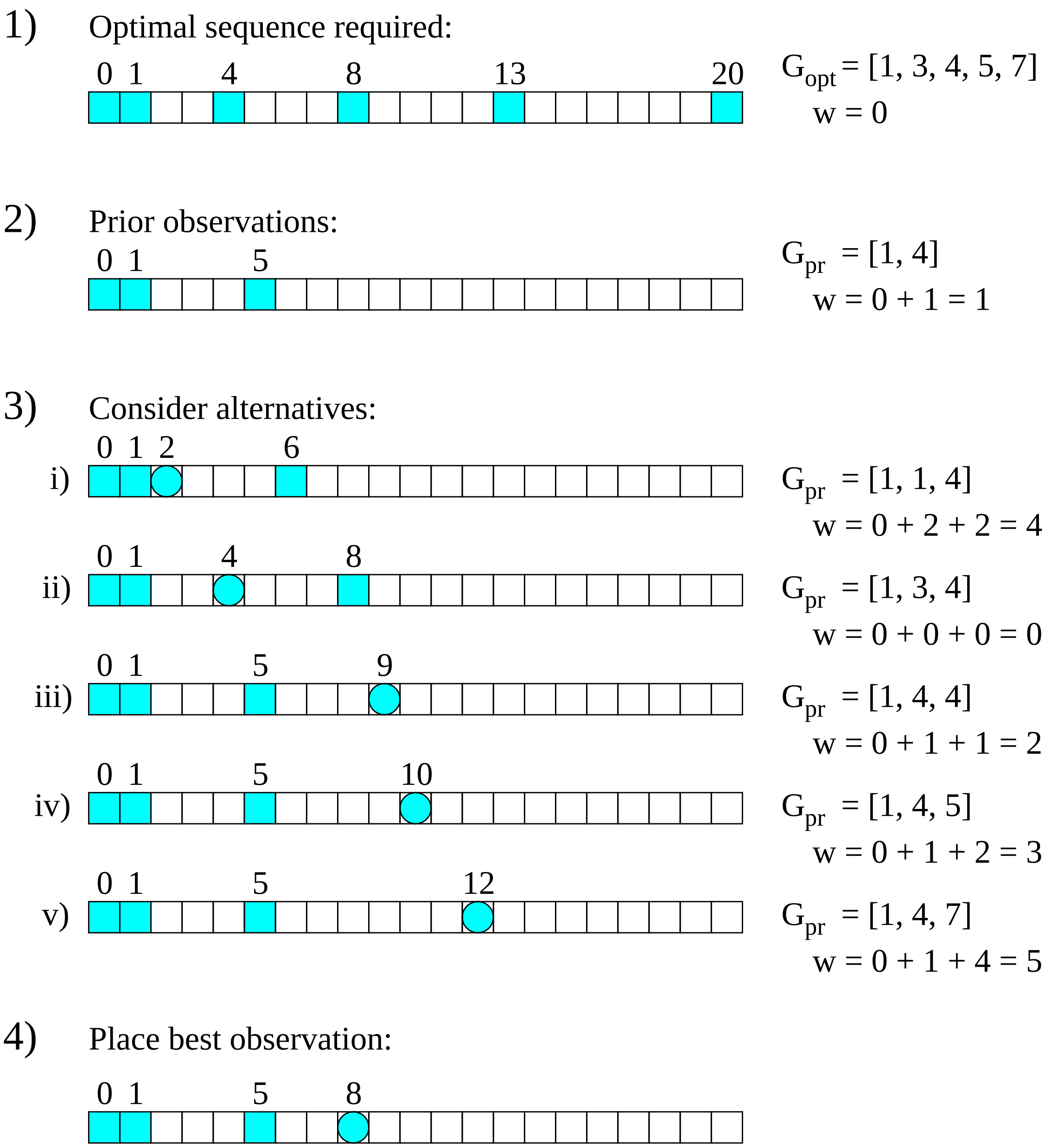

Figure 4.6: A worked example showing how the optimality algorithm described in the text can lead to the selection of gaps other than the shortest under certain circumstances. In step 1, the optimal set of observations is defined. G_{opt} describes the optimal set of gaps. In step 2, the agent calculates the current optimality, based on observations already obtained. In step 3, the possible alternatives allowed by the algorithm are considered in turn. The agent calculates the resulting optimality w if an observation was placed with each of the gaps in the set G_{opt}. It is found that the lowest optimality arises when a gap of 3 is applied (step 3.2). In fact, for this example this choice recovers the optimal sequence (with a perfect optimality of 0). The gap is therefore selected and the new observation placed (step 4).

results leads to a complex quagmire of interactions between the observer and the observations. It is possible, for example, to converge on a spurious signal in the periodogram, and to choose a sampling that selectively reinforces that signal. More seriously, making decisions like this during the collection of the data renders straightforward statistical analysis impossible (Scargle, 1982; Wall & Jenkins, 2003, Ch. 5).

The phase coverage metric S and the interconnectivity metric N are not used either. Although they could be usefully applied to the problem, the specific issues that they address are implicitly handled in other ways. The optimality criterion implicitly deals with phase coverage because it leads to a balanced selection of gaps, which provides relatively even period sensitivity. The choice of run length and number of observations made by the astronomer determines the effective interconnectivity of the time series before the run even begins. These metrics are still useful for evaluating the properties of the run after the fact but, unlike the optimality criterion, are not sufficiently coupled to the definition of the optimal sampling to adequately guide the behaviour of the algorithm on their own.

One idea that was tested but ultimately abandoned was to provide a way for an astronomer to specify a 'redundancy' parameter that would instruct the agent to place observations more frequently than required by the optimal sampling. The rationale behind this was to allow the astronomer to provide some indication of the expected failure rate of the telescope network based on prior experience. Many of the additional observations would be expected to fail, but the total number of observations made in a given time period would be increased.

A lookahead window (a kind of decision horizon beyond which no plans are laid) was implemented to allow the placement of these 'redundant' observations. Additional observations were constrained to lie within the window, to ensure timely feedback when the additional observations was obtained. As more observations were completed successfully, the window was expanded to allow the algorithm more freedom in its choice of timestamps. In this way the algorithm was able to adapt to the actual performance of the telescopes.

Although the idea of redundant observations is conceptually appealing, no evidence was found in any of the tests to indicate that adding such observations improved the performance of the agent. Although the number of observations completed was higher, the overall optimality of the observed series was in every case markedly poorer, regardless of the failure rate of the telescopes in the network. In particular, redundant observations make it impossible to recover a perfect optimal sampling even under ideal conditions.

In order to investigate the trade-off between extra observations and poorer optimality a lightcurve with a fixed relative period of 0.1207143 (corresponding to a 1.69 day period in a 14 day run, a typical but non-integer period) was sampled with different numbers of redundant observations. Both noisy and noiseless lightcurves were generated, and the resulting periodograms compared. The quality of the inspected periodograms was significantly worse for every experiment that made use of redundant observations. Closer analysis of the set of observation spacings showed that the problem was that too many observations were placed with extremely short gaps, leading to poor frequency sampling in the periodogram. This is the crux of the problem: demanding the gaps be larger leads back to the optimal sampling, while demanding more observations makes the gaps smaller, with catastrophic consequences for the periodogram. We thus concluded that redundant observations were not a useful concept.

4.4 Simulations

4.4.1 Implementing a virtual telescope network

The algorithm was tested in two ways. Concepts were initially tested and refined by running the core of the algorithm in a simplified test environment. The progress of time was modelled by a simple loop, and timestamps were generated in the range 0–1 to simplify analysis. This prototyping phase allowed ideas to be tested quickly, with minimal implementation overhead, and modified or discarded based on performance.

Once the main features of the algorithm appeared stable, it was moved to a second, much more sophisticated test environment. This environment was designed to provide a virtual telescope network that would replicate as closely as possible the actual operating conditions under which the agent was to run. This was facilitated by the highly modular web services architecture of the eSTAR codebase. The existing node agent code was modified to replace the backend connections to the underlying telescope with a custom module that would emulate some of the functions of a real telescope. This was possible because the agent-agent interactions are abstracted from the underlying telescope implementation.

The virtual telescope makes use of the `Astro::Telescope` and `Astro::Coords` CPAN modules, written by Tim Jenness[5]. These Perl modules provide a convenient interface to the SLALIB astrometry library (Wallace, 1994). Each instance of the virtual telescope can be 'sited' at a different virtual location. This information allows the virtual telescope to calculate sunrise and sunset times, and to determine the rise and set time of arbitrary points on the celestial sphere. In this way the virtual telescope can accurately calculate whether an object can be observed at a particular time, and use this information to return a score reply. To make the environment more challenging, non-zero score replies have random values between 0 and 1, and have no correlation with the likelihood that the observation will in fact be successful.

The passage of time in the simulation is regulated by the *time server*. This is a simple standalone process that provides the current simulation time via a socket interface. Internally, the time server calculates the current simulation time by scaling the amount of real time elapsed since instantiation by an acceleration factor, provided at startup. Since all agents in the simulation use the time server as the canonical source for timestamps and timing calculations, the time server allows the simulation to be run many times faster than the real world, enabling full simulations of a likely observing run to complete in a reasonable timeframe. Setting the acceleration factor to 1 allows the timeserver to run at normal speed, and therefore provides transparent access to the real-world time.

The probability of observation success is specified at startup for the virtual telescope, and a random number generator used to determine the success or failure of each observation at run-time. For each observation queued at a virtual telescope, that telescope periodically polls the time

[5] `http://search.cpan.org/~tjenness`

server, and compares the current time with the requested start and end times of the observation. If the start time has been exceeded, the observation's status is resolved. If successful, a fake FITS header is generated, with the correct observation timestamp placed within the header. This header is encapsulated inside the RTML message of type 'observation' that is returned. Otherwise, a 'fail' message with no data product is returned. Using the existing user agent and node agent codebase meant that at the protocol level the interaction between eSTAR components in the simulator was identical to that in the real world, allowing most aspects of document exchange to be tested before live deployment.

One limitation of the simulation environment was that fail rates at the virtual telescopes were not correlated, that is, periods of failure were not consecutive in time. This was not implemented because of the complexity of choosing realistic correlation behaviour and subsequently inferring meaningful results from the simulations. Even without such behaviour, the most critical aspects of the agent were able to be adequately tested. Nevertheless, this is an obvious way in which the virtual telescope environment could be improved if accurate performance measurements are desired.

4.4.2 Simulation results

The purpose of the tests performed in simulation was to identify bugs at several levels, and to evaluate the performance of the algorithm and iterate improvements in a tight feedback cycle. This included analysis at the message-passing level for conformance with the HTN protocol specification. The behaviour of the implemented algorithm was carefully compared with the design statement. A number of discrepancies were found and corrected. Most importantly, the performance of the algorithm under pseudo-realistic conditions allowed a number of unforeseen corner cases to be identified and correctly handled.

The scheduling algorithm was tested in a number of simulated observing runs, but the discussion of a single experiment is sufficient to illustrate the general performance characteristics. In this simulation a 10 day, 60 observation run was attempted. An acceleration factor of 50, applied to the time server, allowed a single 30 minute observing window to complete in 36 s. The start of the run was set to a date of 27/04/07, at which time the target was observable approximately 2/3 of the time from at least one of the three Robonet-1.0 virtual telescopes. In this simulation, in addition to the non-observable blocks (which returned a score of 0 to the agent), the virtual telescopes were set to randomly fail observations at run-time with 50% probability (an arbitrary fraction that seemed reasonable).

The experiment was halted after 10 days had elapsed in simulation time. It was found that of all the observations submitted to the network, 40 had succeeded (i.e. been accepted to the queue, and then been declared successful at some point in the specified observing window), while 50 of the queued observations had failed (queued but not observed). All the successful observations were found to have close-to-optimal timestamps, subject to the observability and success constraints.

The results indicated crudely that if the real network exhibited a similar fail rate, then a ballpark figure of approximately 2/3 of the observations to be completed successfully was a reasonable expectation by the 10th continuous day of observing, and additionally, that the spacings

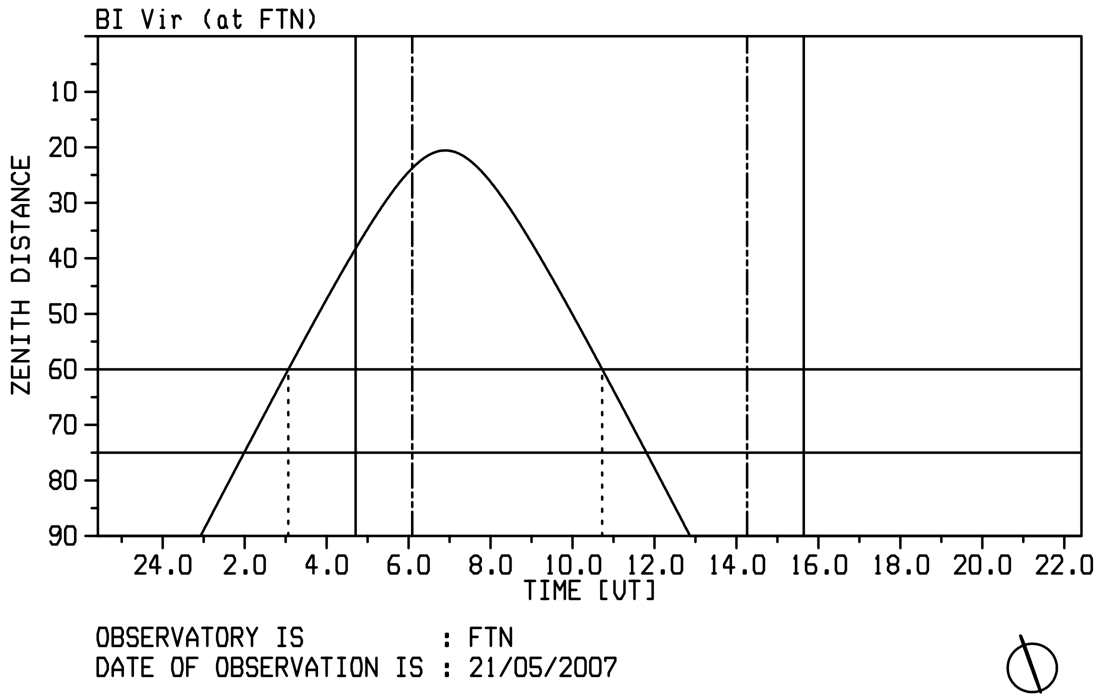

Figure 4.7: BI Vir ephemeris from the Faulkes Telescope North, Hawaii at the start of the observing run.

between individual observations were likely to be near-optimal. As previously mentioned, the chief limitation of the tests was that fail rates at the virtual telescopes were not correlated, a feature which was found to significantly complicate performance evaluation in the real observing run, as we shall see in the next section.

4.5 Observing with the adaptive agent

4.5.1 Observation parameters

The RR Lyrae star BI Vir, with coordinates α = 12 29 30.42, δ = +00 13 27.8 (J2000, Henden & Stone, 1998), was chosen as the target of the adaptive scheduling programme, both for its visibility from all three Robonet-1.0 telescope sites and because it has a short, well-defined period of 0.3357 days (8.056 hrs). Additionally, at an R-band magnitude of 14.2–15 it is relatively bright, allowing adequate signal-to-noise to be achieved with relatively short exposures of only 5 s. Figures 4.7–4.9 indicate the visibility of the target from each site at the start of the run. These observing windows are summarised in Table 4.1. The target observability over the course of the run is shown in Figure 4.10. The observability is defined here as the length of time the target was above a horizon of 30 degrees while it was night (bounded by astronomical twilight).

To create realistic period search constraints, the following parameters were adopted. To ensure a significantly undersampled dataset, $\nu/\nu_{\rm N}$ was fixed at 3. The calculation was based on a run length of 5 days, to provide sensitivity to periods roughly an order of magnitude above the

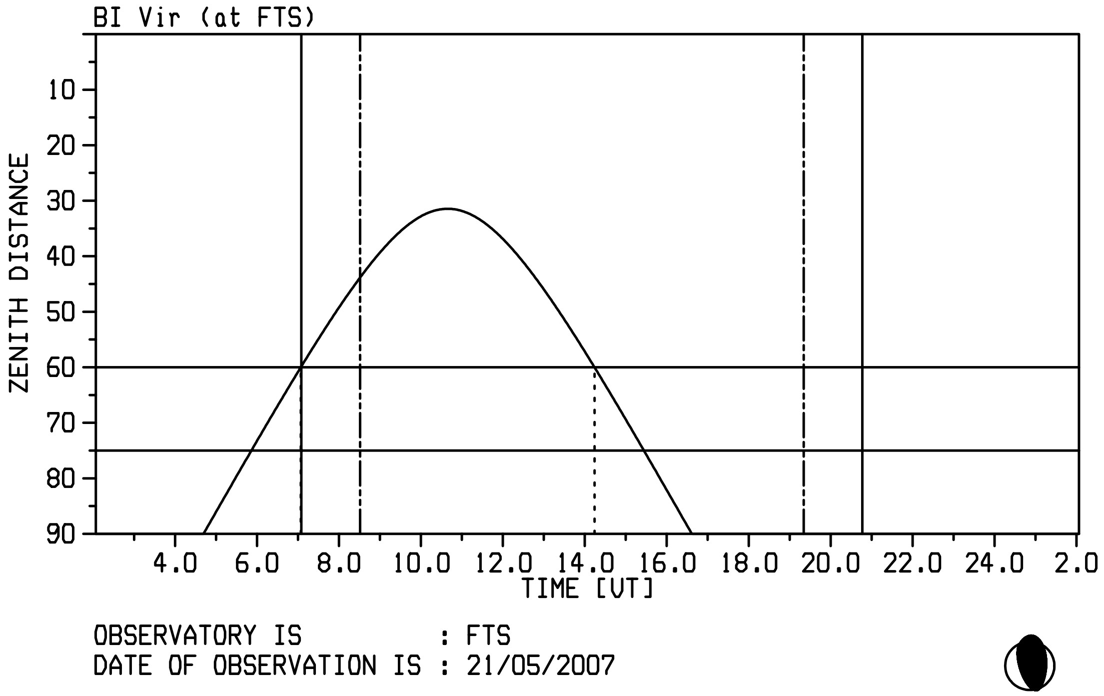

Figure 4.8: BI Vir ephemeris from the Faulkes Telescopes South at Siding Springs, Australia at the start of the observing run.

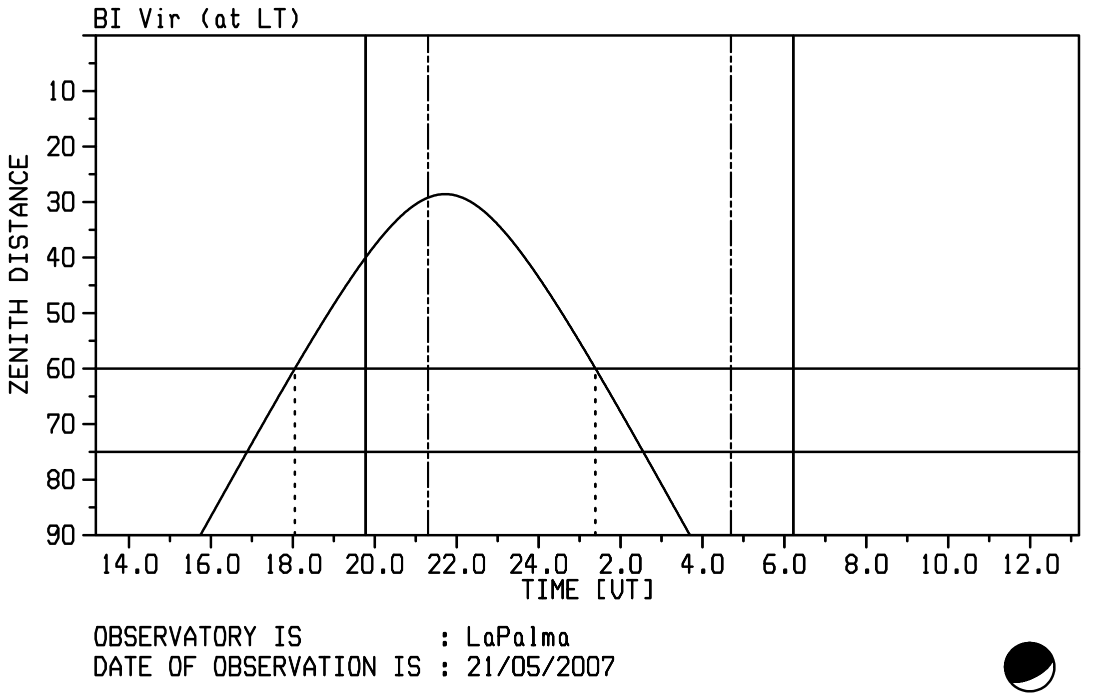

Figure 4.9: BI Vir ephemeris from the Liverpool Telescope on La Palma at the start of the observing run.

Telescope	First Observable (UT)	Last Observable (UT)	Duration
Faulkes North	06:00	10:45	4 hr 45 m
Faulkes South	08:30	14:15	5 hr 45 m
Liverpool	21:15	1:30	4 hr 15 m

Table 4.1: Observing windows for BI Vir from the three Robonet-1.0 telescopes, at the start of the observing run (21/05/2007).

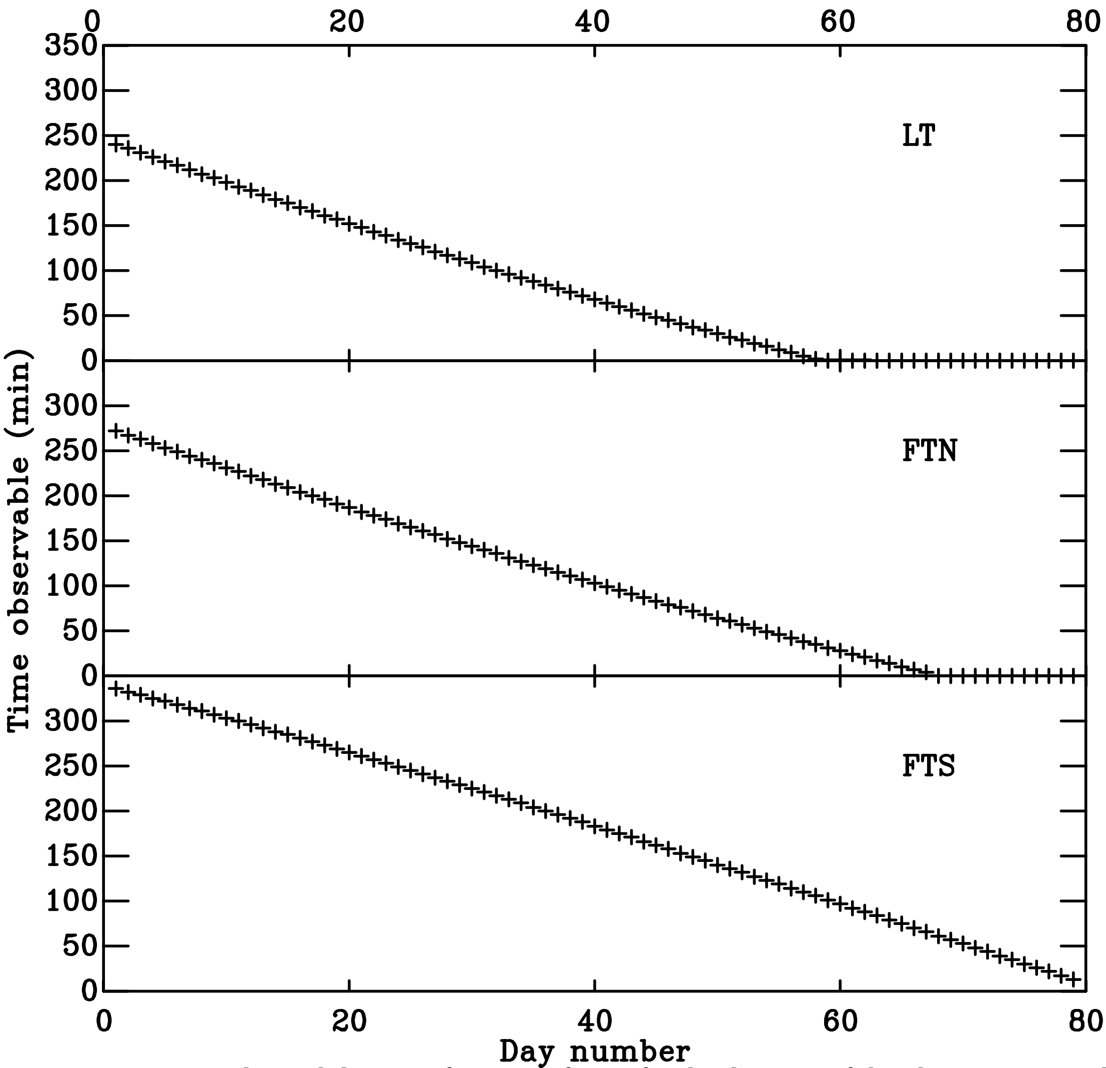

Figure 4.10: Target observability as a function of time for the duration of the observing run. The observability is defined as the length of time the target was above a horizon of 30 degrees while it was night (bounded by astronomical twilight). After 40 days, the target was observable on average for just under two hours per telescope per night. By 60 days, this had fallen to an average of just over an hour per telescope per night.

true period, after taking into account likely overrun from failed observations. Sixty observations was chosen as a reasonable minimum number of observations for a useful dataset, in order to keep the total telescope time used for the experiment small. Applying Eqn. 3.16 then gives

$$\begin{aligned} \left(\frac{\nu}{\nu_{\mathrm{N}}}\right) &= \frac{2\nu_{rel}}{N} \\ 3 &= \frac{2\nu_{\mathrm{rel}}}{60} \\ \nu_{\mathrm{rel}} &= 90 \end{aligned}$$

which is equivalent to a minimum relative period of 0.011. For a total run length of 5 days, this means the sampling should be sensitive to periods in the range 1.33 hrs $< p <$ 2.5 days, providing good period coverage around the true period of 8.056 hrs.

Factors outside of the adaptive scheduling agent's control such as telescope load can obviously impact the experiment. To negate the effect of such variations on the results, a control programme was run concurrently with the ASA programme to allow the effect of external, telescope-related parameters to be discounted. In order to make the test as rigorous as possible, the control observations were placed in a manner typical of an astronomer without access to an adaptive agent, and instead leveraged the telescope scheduler's own observation constraint model. The control requests were monitor group observations, each requesting two observations, with an interval of 4.5 hours and a tolerance of 2.25 hours. These were placed twice daily at fixed times using a simple script. This submission strategy is typical of an astronomer seeking a reasonably constant cadence over the duration of the run. The relatively high tolerance means that the second observation in each request may be placed anywhere between 2.25 and 6.75 hrs of the first observation, providing significant freedom to the telescope scheduler. The existence of a fairness metric used by the scheduler to moderate the chance of selection for observation was a potential concern, because it could mean that the successful placement of control observations would adversely affect the chance of success of an ASA observation. However, the fairness metric operates on a semester timescale, so this was not considered to be an issue (S. Fraser, priv. comm., 05/07).

4.5.2 Results

The target had originally been chosen with the expectation that the run would begin in mid-April 2007, at which point BI Vir would be close to the zenith. The first test observation was in fact obtained on 30/04/07, but a wide variety of teething problems, as well as external commitments (including the requirement that the author attend his own wedding!) meant the system was not considered stable until 21/05/07. This date is therefore taken as the 'official day 0' of the run in what follows, and any earlier observations were discounted.

Halt			Restart		
Date	Time	Explanation	Date	Time	Downtime duration
21/05		[ASA observations begin]			
28/05	00:47	Memory leak	29/05	18.49	18 hrs 2 m
29/05		[Control observations begin]			
03/06	05:32	Power cut	05/06	15.59	2 d 10 hrs 27 m
11/06	19:38	Memory leak	12/06	03:16	5 hr 38 m
12/06	11:18	Idling + memory leak + travel	21/06	16:43	9 d 5 hr 25 m
22/06	11:19	One month boundary bug	22/06	15:25	4 hr 6 m
28/06	12:48	Pre-emptive restart	28/06	13:02	14 m
05/07	14:05	Pre-emptive restart	05/07	14:08	3 m
07/07	01:50	No telescope return document	09/07	12:22	2 d 10 hr 32 m
16/07	02:38	Memory leak	16/07	13:31	10 hr 53 m
17/07	12:55	No telescope return document	17/07	12:59	4 m
23/07	10:25	Memory leak (hung since 22/07)	23/07	10:31	16 hr 31 m (effective)
25/07	21:11	[Last successful observation]			
31/07	02:36	Memory leak	31/07	14:10	11 hr 34 m
07/08	06:41	[Final halt]			
			Total downtime:		16.3 d

Table 4.2: Log of agent halts and restarts. The majority of the downtime arises from a single event (12/06). In some cases the downtime duration is very small - this indicates a manual shutdown and restart.

Downtime

For a number of reasons the agent was incapacitated at various points during the run, and was unable to submit observations for some time. These were in some cases due to a lack of protective error-handling code. For example, occasionally errors at a telescope would cause it to default on the prescribed return of a completion document ('observation' or 'fail') to the agent. This locked the agent into a permanent sleep state, because no timeout sanity checking had been implemented. Another recurring problem was a persistent, slow memory leak which lead to a system crash approximately every 10 days. The agent could be correctly reinitialised without loss of information because of the serialisation mechanism by which all data are logged, but there were occasional delays due to the lack of a human operator to initiate the restart sequence. Other downtimes were caused by network outages and in one case a power cut. Table 4.2 shows the full details.

Altogether the total downtime of the agent was 16.3 days, or approximately 20% of the total run. The majority of this is accounted for by a 9 day outage which was caused by a combination of an idling bug and a fail state which was not detected for several days, due to travel by the author. Additionally, a series of bugs in the script for automated submission of the control observations meant that there was a delay of approximately 7 days from the start of the ASA run until the first successfully submitted control observation.

Placement of observations

In all, 33 ASA observations were successful, out of a total of 250 queued (recall that a successfully queued observation indicates a positive score from at least the telescope that has accepted the observation into the queue). This is a success rate of 13.2%. In contrast, the control series achieved only 26 observations, but the total number queued was significantly lower, at only 69, yielding a success rate of 37.7%.

These results indicate two things. In the first place, the adaptive scheduling agent's strategy of speculative observation placement, always accepting any non-zero score, leads to a greater rate of failure, but produces a raw increase in the total number of successful observations of almost 25% compared to the control strategy. On the other hand, the higher success rate of the control strategy indicates that when the telescope scheduler itself decides to queue an observation, that observation is much more likely to actually be successful. This second result is perhaps unsurprising. By design, the degree of speculation exhibited by the agent was deliberately maximised, because of the threat of non-completion of the full number of observations. A more cautious agent could be obtained by increasing the acceptable score threshold. This would be equivalent to placing more 'trust' in the telescope's ability to judge its own observing load.

On the other hand, the fact we can obtain an increase in successful observations of more than 25% via a scheduling agent is quite startling. The implication is that the decision-making of the telescope scheduler is rather more conservative than that of the agent. This illustrates the primary advantage that the eSTAR system gains by working at the level of the smallest schedulable atomic block — a single exposure. The subtlety is that the concept of groups for the robotic telescope scheduler is more than just a way of linking observations. Since the scheduler does not know what fraction of the completed group is an acceptable outcome for the user, it is forced to err on the side of caution, and delay obtaining observations unless it is reasonably sure that the whole group will succeed. This has the effect of simplifying the decision-making process, because the size of committed blocks is much larger. Working at the level of individual observations, the adaptive scheduling agent takes on the responsibility for deciding whether individual observations are worthwhile. Since the agent can dynamically adapt its ongoing strategy, it is possible to achieve much finer-grained control than the telescope scheduler itself, but the cost is the extra complexity which must be handled by the algorithm in real-time.

This result clearly indicates that from a user perspective it is much more efficient to take on the extra burden of micro-managing the observing plan, if one possesses the ability to do so. Regular cadencing and group specifications are conveniences that help keep life tractable for the astronomer, and mitigate against the uncertainty of the actual observing run. But they are gross simplifications when compared to the ideal specification of the plan, one that can be adapted dynamically in response to changing circumstances.

Be that as it may, this is secondary to the main result. For the purposes of this problem, the critical point is the size of the gaps in time that have been created between individual observations. Figure 4.11 indicates the positions of successful observations with respect to the run overall. Figure 4.12 does the same for the set of control observations. Both runs indicate a dearth of observations between 16 and 32 days. This common behaviour across both datasets indicates a lack

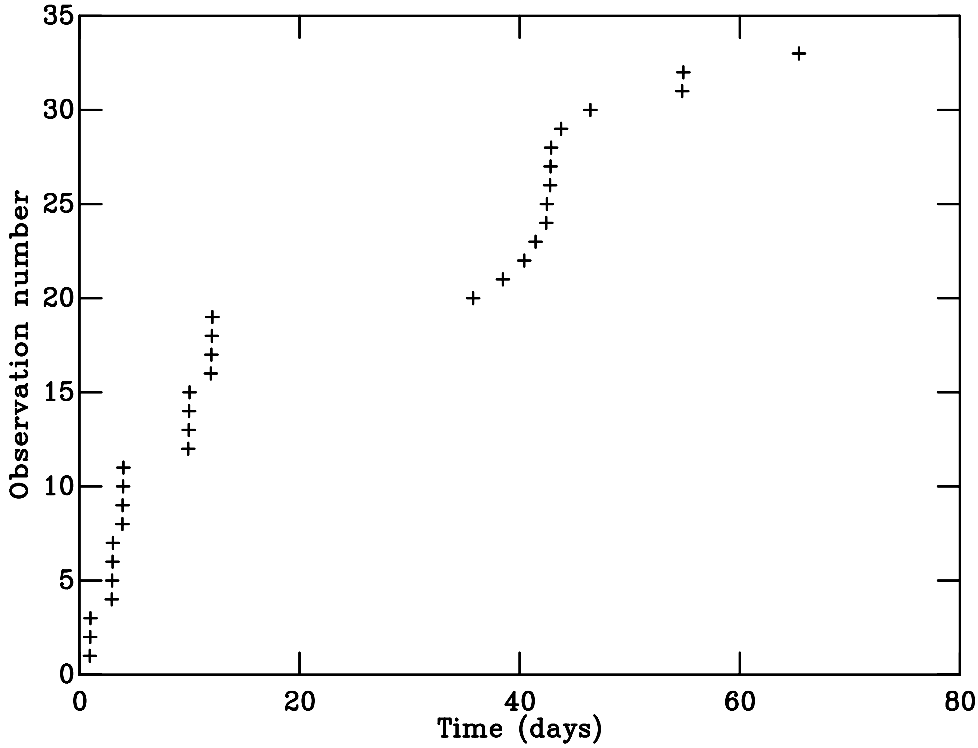

Figure 4.11: The distribution of ASA observation timestamps over the course of the observing run (33 individual frames).

of resource, which is a function of available on-sky time and the quantity of competing requests, among other variables. Both runs peter out at around 60 days. At this point in the run there is an average window of less than an hour per telescope per night in which to observe the target (see Fig. 4.10, above, for more details).

Figures 4.13–4.14 plot the size of the gaps between concurrent observations obtained by the ASA. In Fig. 4.13, the gaps are indicated in the order in which they were acquired. This is a combination of gaps requested by the agent, and other gaps created by the absence of observations. To give a sense of scale, the optimal set of gaps is overplotted. It can be seen immediately that many of the gaps are much larger than desired for optimal sampling. These indicate large breaks in coverage where no telescope on the network was willing or able to obtain the interim observations necessary to keep the gaps small enough to be useful.

Taking advantage of the reordering property of the gaps clarifies the picture. Figure 4.14 shows the same data, but this time with the observed gaps ordered from shortest to largest. Around half of all observed gaps lie on the line of optimal sampling. More specifically, the agent has managed to obtain the shortest third of the required set of gaps. The remaining gaps are almost all much larger than required.

This result vividly demonstrates the algorithm's successful implementation of the optimal sampling strategy. Whenever possible, the agent has tried to fill a required gap. The agent has concentrated on the shortest gaps, because they are most important, given that (overly) large gaps are inevitably being obtained as a result of telescope non-cooperation. At every stage in the run, the agent will adapt to improve its circumstances, if it can. If, for, example it was able to acquire

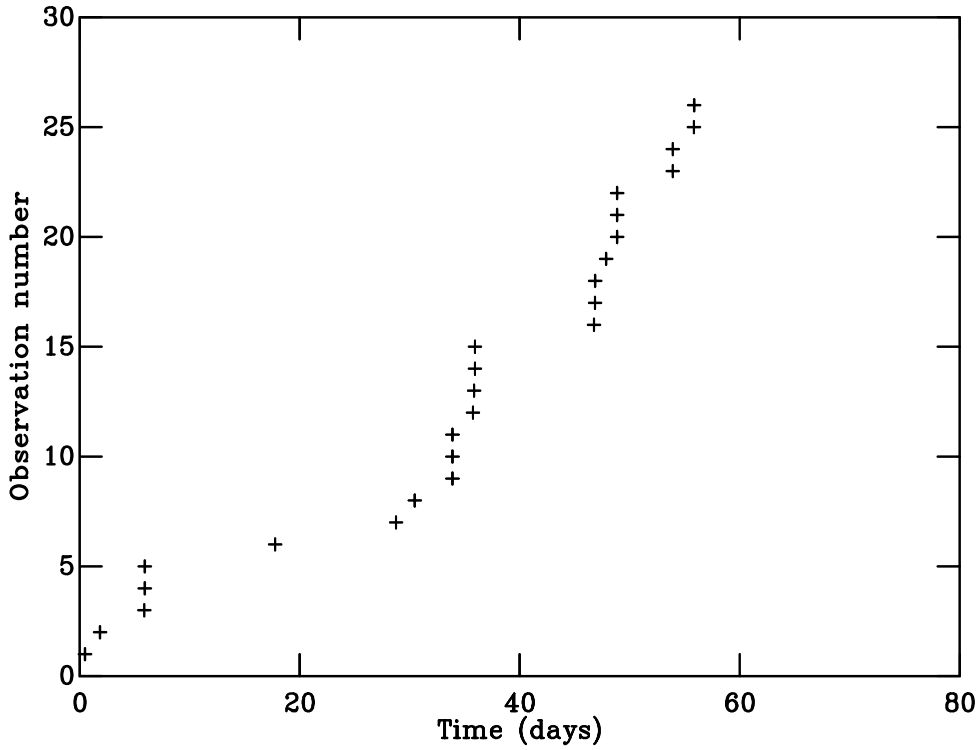

Figure 4.12: The distribution of control observation timestamps over the course of the observing run (26 individual frames).

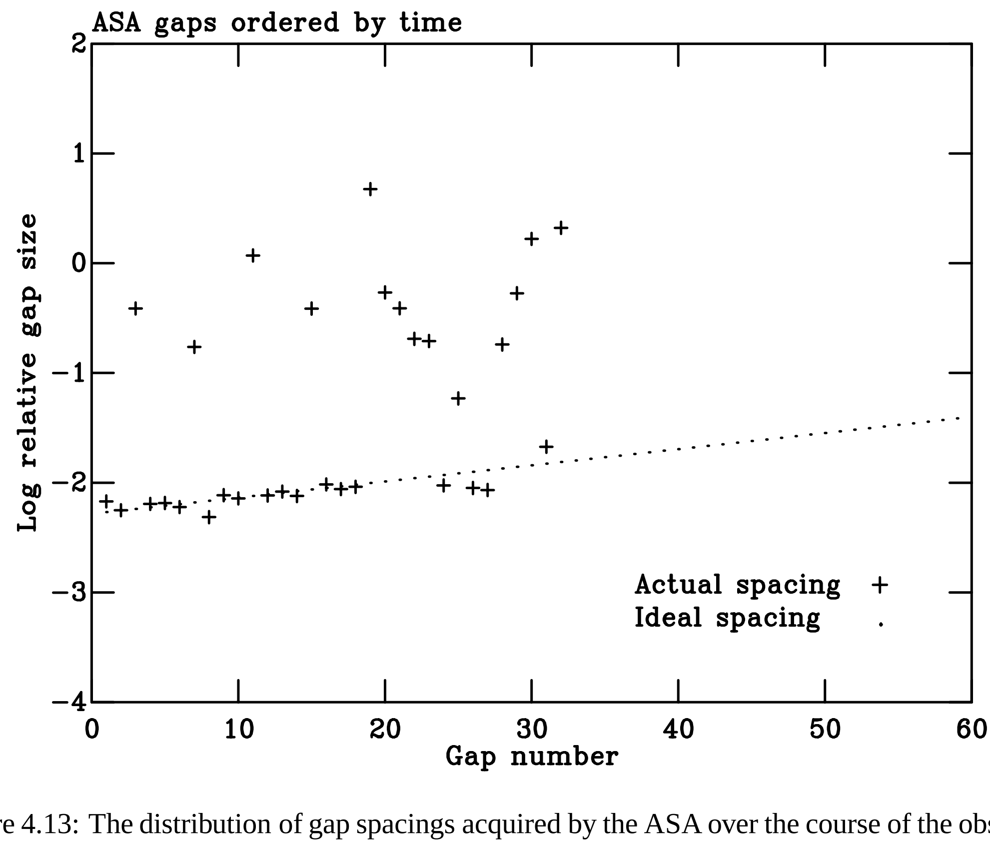

Figure 4.13: The distribution of gap spacings acquired by the ASA over the course of the observing run. The ordinate axis is the logarithm of the relative gap size, i.e. the gap as a fraction of the total run. The set of ideal optimal spacings has been overplotted for scale.

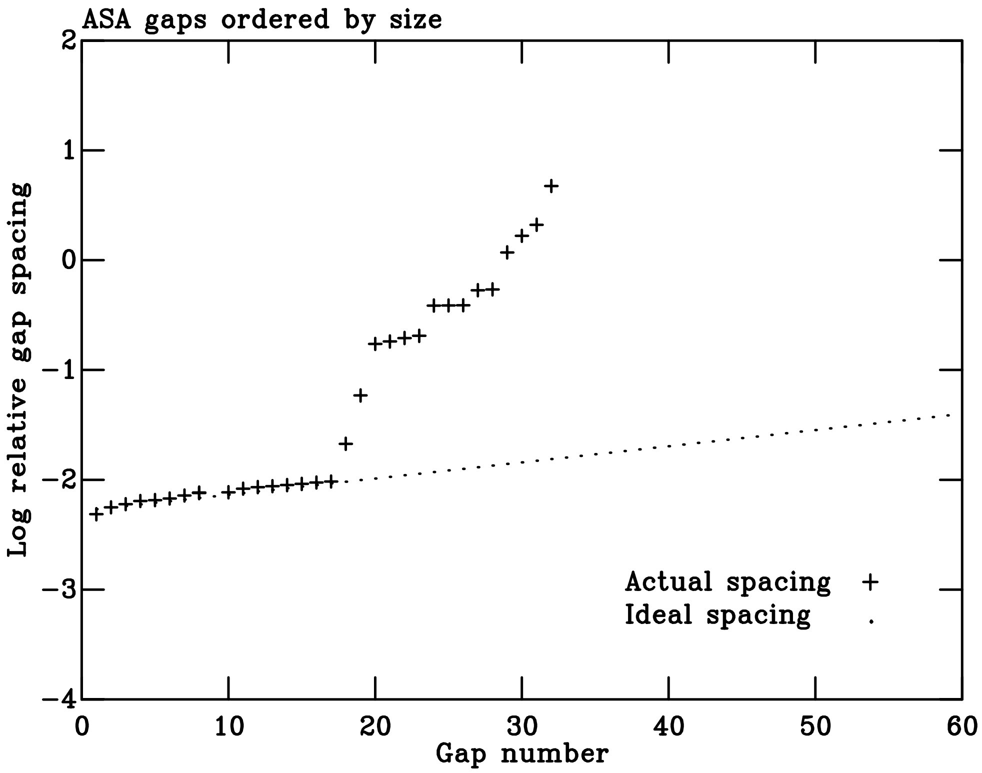

Figure 4.14: Comparison of the actual gap spacings acquired by the ASA against the ideal optimal spacing. For ease of comparison the acquired observations have been ordered by size.

the final 27 observations in ideal locations, the effect would be to shunt all the larger gaps to end of the graph. Thus poor spacing positions early on in the run need not dictate the ultimate fate of the run.

The 10th gap was not obtained by the ASA. This is because the agent received an 'observation' message from the telescope, indicating that this observation had been successful, but during the subsequent data processing stage the telescope ultimately chose to reject the observation, on the grounds of poor focus. Since the agent was unaware that the observation had effectively failed, it continued on the assumption that it had been successful. This highlights a shortcoming of the current system, namely that post-processing of observations can invalidate the message previously sent to the agent.

The existing run is partial in the sense that only 33 out of the requested 60 observations were actually obtained. This is a typical risk associated with the robotic mode of observing. The agent has no knowledge of whether the number of observations it has obtained is adequate for the science needs of its user. However, it can guarantee that such observations as there are have been placed optimally, as far as that was possible. Therefore the partial dataset is the best sampling that could have been obtained under the actual observing conditions experienced during the run.

For comparison, Figures 4.15 and 4.16 present the equivalent plots for the control run. The difference in gap quality is abundantly clear. Like the ASA run, the control run inevitably also suffers from the problem of overlarge gaps, but without the focus on achieving optimal placement, those gaps that are of use are quite arbitrarily and wastefully positioned. This is a common consequence of regular cadence observing requests. While the data may still be useful, the quality in

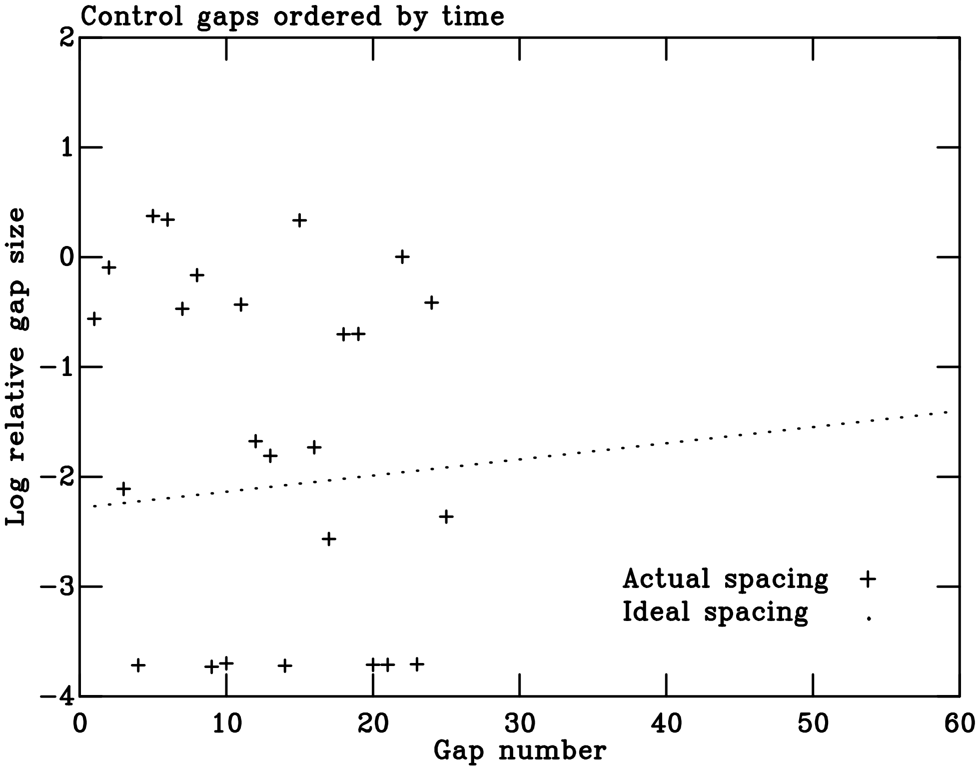

Figure 4.15: The distribution of gap spacings acquired by the control programme over the course of the observing run. The set of ideal optimal spacings has been overplotted for scale.

terms of period coverage must inevitably be poorer. It is important to note that this would be true even if the control run had managed to acquire the same number of observations as the ASA run.

Both runs overran by many days, and achieved only around half of the observations that were expected. A number of factors contributed to this lack of observation success. As previously discussed, the delayed start to the run meant the target was already past the zenith at local midnight, so the total number of hours of potential on-sky time was significantly less than the ideal, and this fraction continued to drop (see Fig. 4.10) with time. A second significant issue was a period of prolonged poor weather at Faulkes South. Finally, the periods of agent downtime also extended the run beyond the initial estimate. Although this was bad from the perspective of the science programme, it gives some idea of the wide range of factors that can and do contribute to the network behaviour in the real world. From the perspective of the experimental evaluation of the agent, this was a useful feature of the observing run. This kind of unpredictable and hostile environment is exactly the type of scenario the agent methodology seeks to mitigate, and a demonstration of this kind is perhaps the only convincing way to test the success of this approach.

Lightcurve analysis

The data were reduced following a similar process to that described in Chapter 2. Optimal extraction was performed to obtain photometry for the stars in each frame. Uncertainties on the flux values thus derived were calculated as described in (Naylor, 1998). To allow the datasets to be combined, the flux for the target star BI Vir was normalised by dividing by the star at

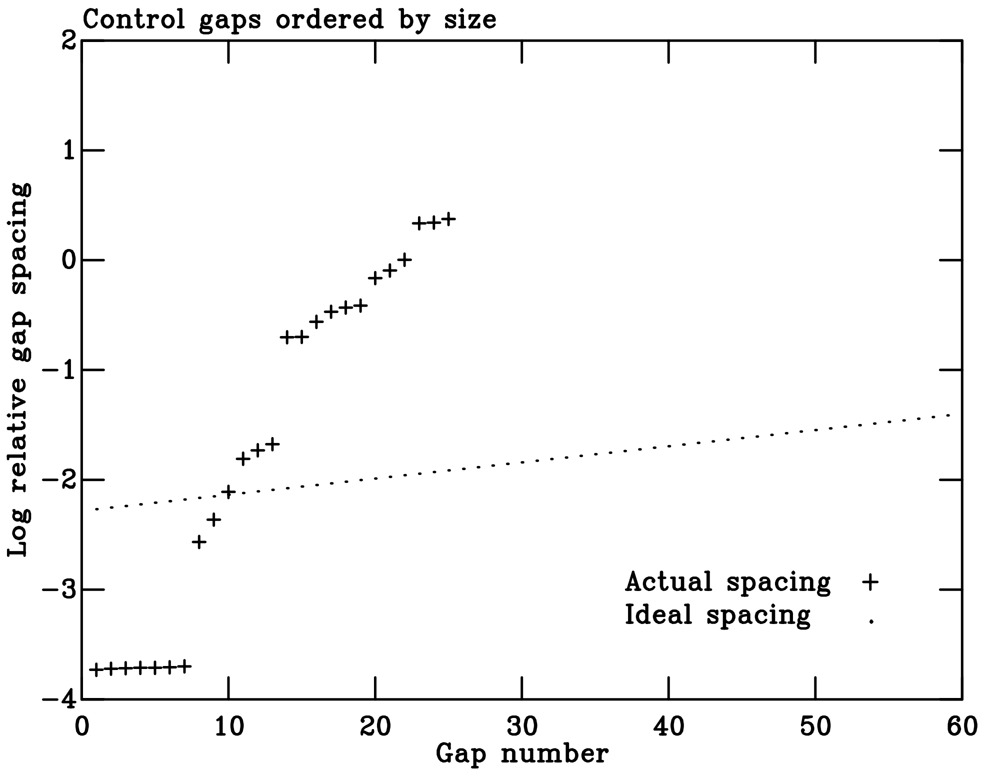

Figure 4.16: Comparison of the actual gap spacings acquired by the control programme against the ideal optimal spacing. For ease of comparison the acquired observations have been ordered by size.

	LT	FTN	FTS	Total
ASA run	7	6	20	33
Control run	2	4	20	26

Table 4.3: Final breakdown by telescope of the useful observations available from each run, after data reduction.

α = 12 29 30.66, δ = +00 14 10.9 (marked '2' in Fig. 4.17) in each frame. Finally, observation time-stamps were corrected to barycentric time. Observations from all three telescopes were combined to produce separate lightcurves for the ASA-driven and control runs. The final breakdown of useful frames acquired from each telescope for each of the runs is shown in Table 4.3.

Figures 4.18 and 4.19 present the phase-folded lightcurves for the ASA and control runs respectively. Although the number of observations in each run was similar, it is immediately apparent that the phase coverage of the lightcurve achieved by the adaptive agent is much better than that of the control run. This distinction is explicitly quantified by the S metric. Recalling that perfect phase coverage is indicated by $S = 1$ (independent of the number of datapoints), operating on the phase-folded lightcurves yields $S_{\mathrm{ASA}} = 2.03$ and $S_{\mathrm{control}} = 3.05$. The root cause of the performance deficit observed with the control run is this large gap in phase coverage. As discussed in Chapter 3, the presence of many similar small intervals is prohibitive to period sensitivity at longer time-scales, and also makes the occurrence of larger gaps in the phase coverage more likely.

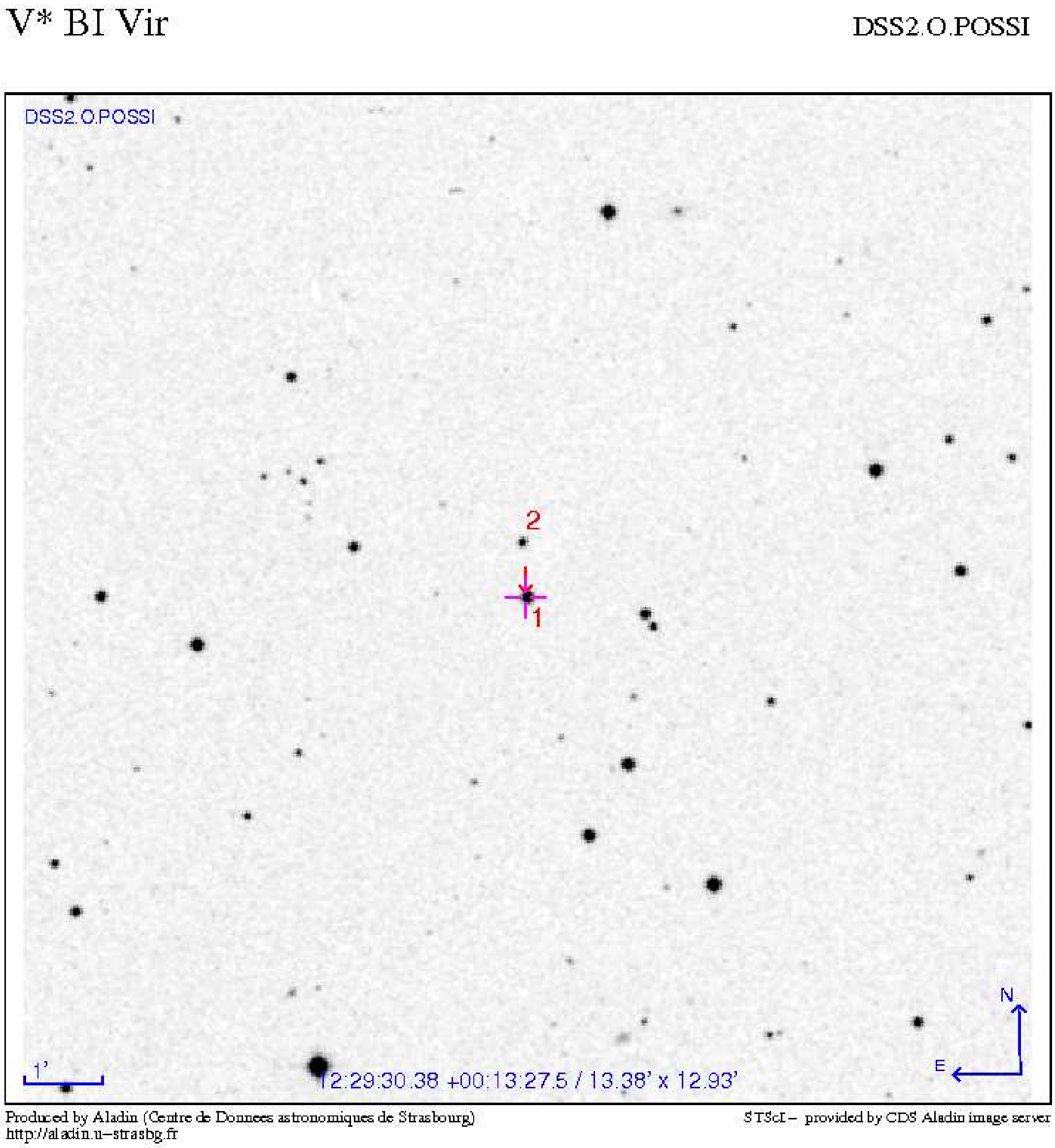

Figure 4.17: The field surrounding the variable star BI Vir (marked '1'). To compare observations between the three telescopes, the flux was normalised by a reference star in the same frame (marked '2').

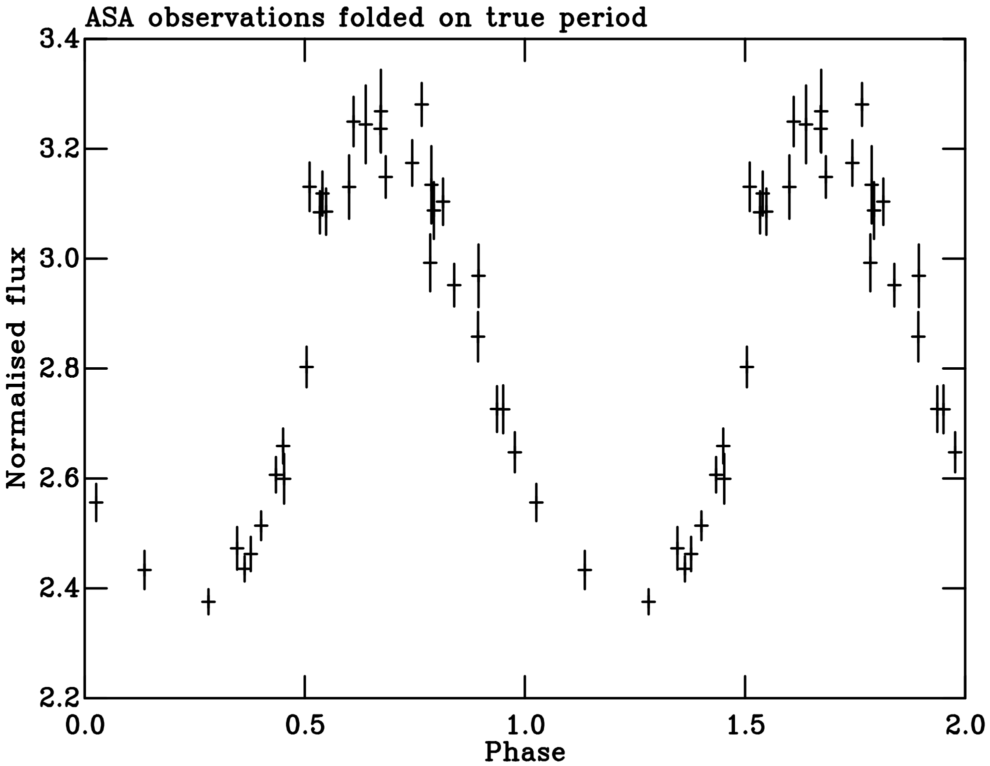

Figure 4.18: The ASA dataset folded on the true period, P = 0.3356548 days.

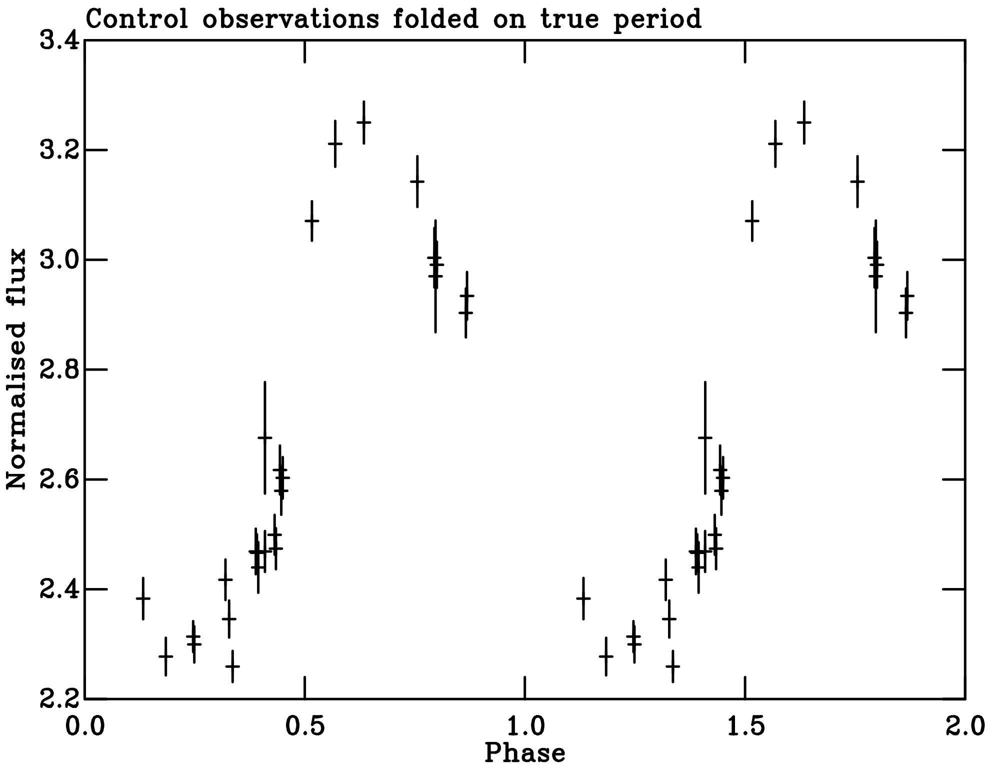

Figure 4.19: The control dataset folded on the true period, P = 0.3356548 days.

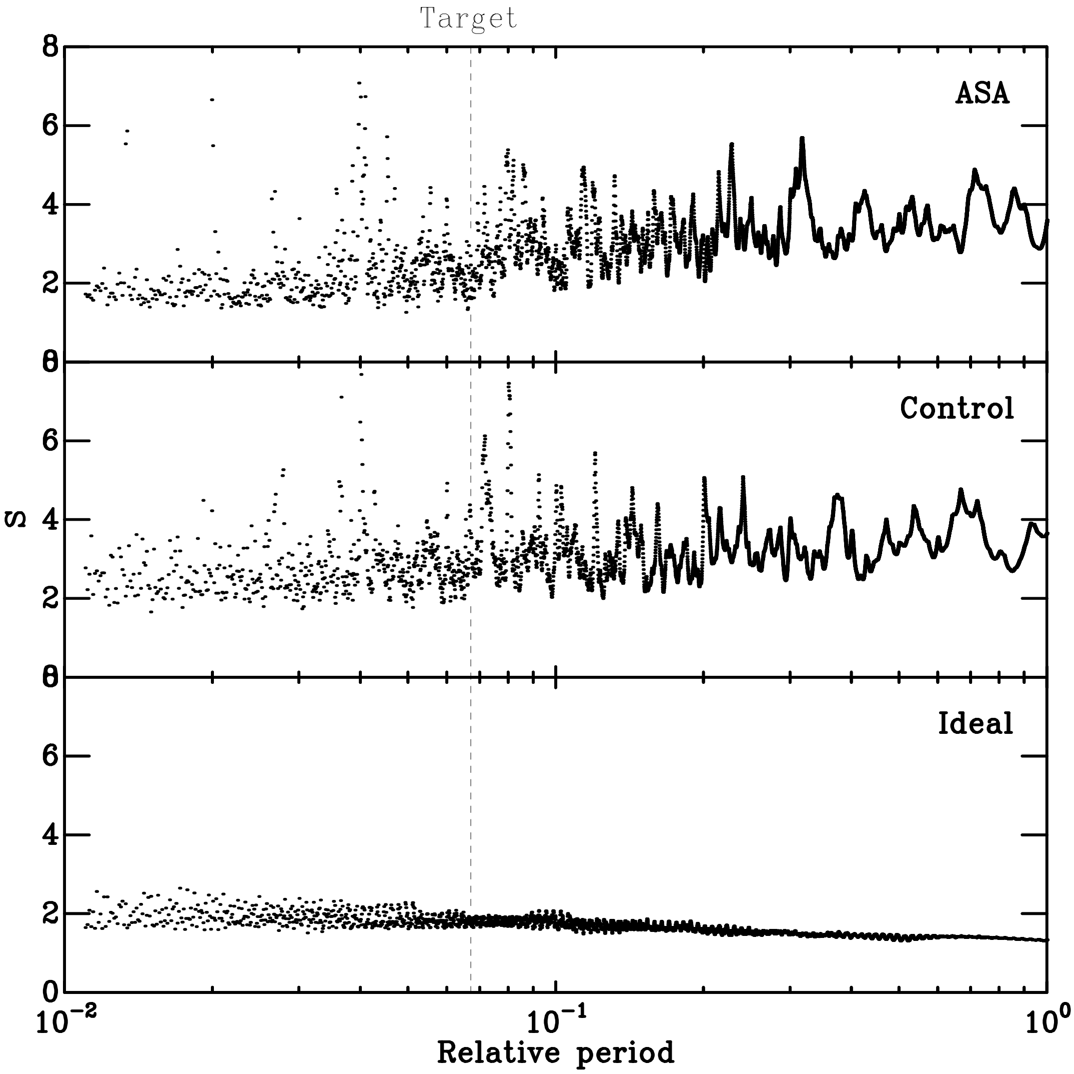

Figure 4.20: Phase coverage plotted as a function of relative period for the two observing runs, showing the relative sensitivity of the ASA (*top*) and control (*middle*) runs to the period range of interest. For comparison, the phase coverage provided by the ideal optimal geometric sampling is also presented (*bottom*). The vertical dashed line marks the true period of the target.

The sensitivity of each run to the period range as a whole is quantified in Figure 4.20, which gives the phase coverage S as a function of period, from a minimum relative period of 0.011 up to the target run length (see Section 4.5.1 for details on how this minimum period was chosen). For comparison, the phase coverage provided by the correct geometric sampling, that sampling which the adaptive scheduling agent was aiming to achieve, is also illustrated. Both runs have substantially poorer period sensitivity than the ideal dataset, as expected. This is partly a consequence of misplaced observations, but a more substantial problem is the general lack of observations — each of the runs achieved only around half of the number of data points present in the ideal sample. Over the full range, the two runs have broadly similar phase coverage properties. Both exhibit undesirable spikes at a number of periods, and coverage of shorter periods is in general better than for longer ones. The mean value of S for periods up to 0.5 is 3.194 to four significant figures, effectively identical for both runs. At short periods, however, the ASA run has substantially better coverage. For periods below 0.1, the mean S values are 3.009 and 2.556 for the ASA and control runs, respectively.

For completeness, Fig. 4.21 plots the interconnectivity of the two runs as a function of period. As before, the ideal geometric sampling is shown for comparison. This figure is not as easy to interpret as the corresponding phase coverage diagram (Fig. 4.20). Both runs have more interleaved cycles than the theoretical sampling, which is an obvious consequence of the much longer actual run length, which was a factor 12 larger than the target run length. Interestingly, although the ASA run appears from casual inspection to have a much greater average interconnectivity than the control run, analysis of the data shows that the average values for the ASA and control run are 83.6 and 62.1, respectively, a much smaller difference than might be assumed. The standard deviation of the ASA interconnectivity is much higher, however, at 30.0%, almost double the control run standard deviation of 18.0%.

The much more uniform interconnectivity of the control run implies that for many periods, phase folding does not split up consecutive observations. The lower spread to the points indicates that the average separation in time between consecutive observations is small compared to a typical folding period. From inspection of the gaps between timestamps this is seen to be the case (see Figure 4.16, discussed previously). Not much else can be directly inferred from the plot. If the runs had been the same length as the ideal sampling (5 days), then the relative difference between each run and the ideal sampling would tell us broadly whether the distribution of gaps was of approximately the right size. If the gaps are on average smaller than the ideal sampling, then the interconnectivity would be generally lower. A high interconnectivity on the other hand would indicate the presence of too many large gaps.

Why is it that so many of the control observations occur so close together? The answer lies with the logic of the telescope scheduler. Control observations were submitted with relatively large, flexible windows, and it was left to the scheduler to duly decide when the observations should be executed. Although the observation requests were specified to be 4.5 hours apart, each one was given a flexible window of 2.25 hours within which the observation could be obtained. This meant that it was possible for the scheduler to place both observations close together, near the intersection point of both windows. If conditions are such that a control observation has been

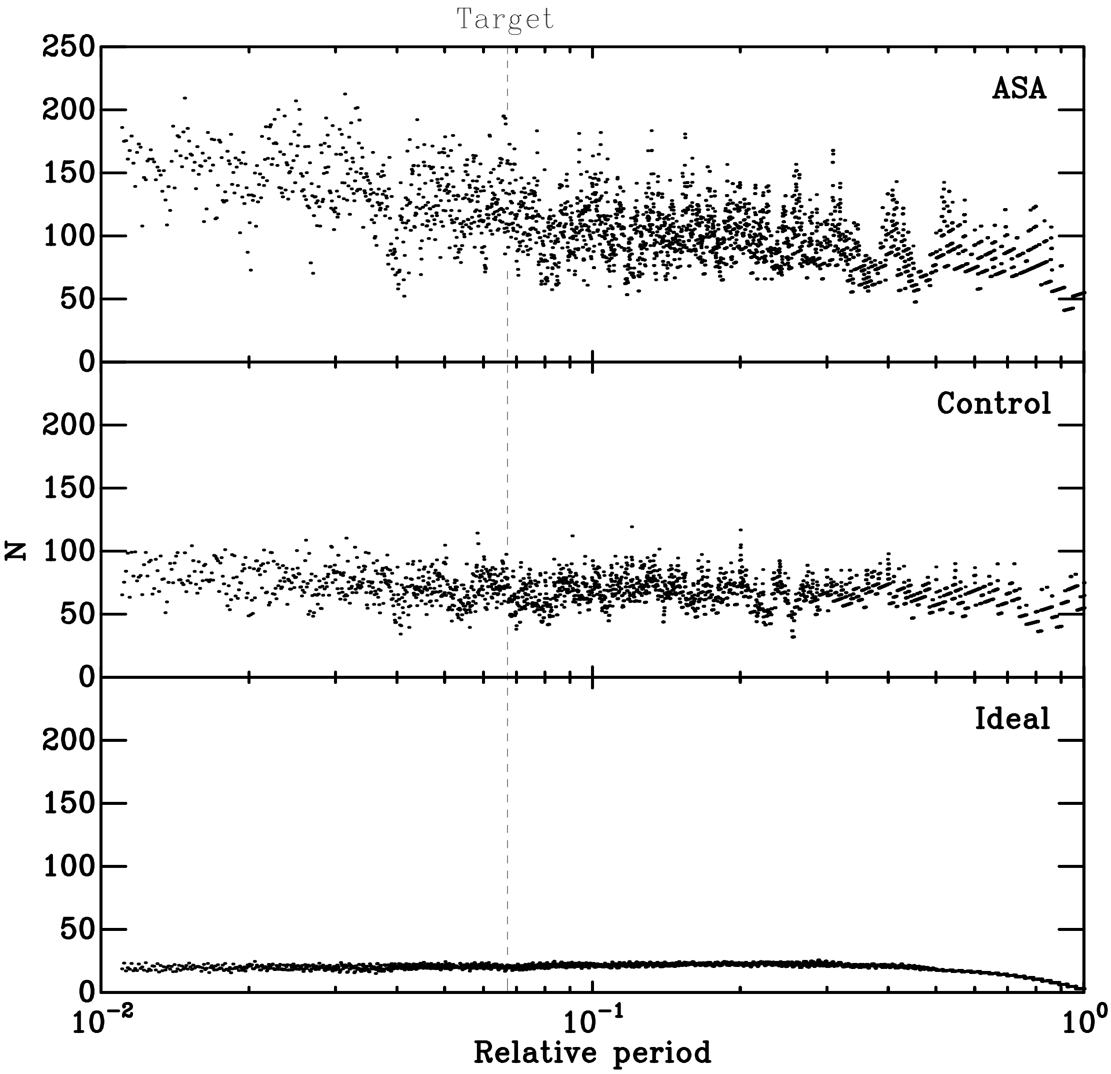

Figure 4.21: Interconnectivity plotted as a function of relative period for the two observing runs, giving an indication of the number of cycles of data being folded at each period for the ASA (*top*) and control (*middle*) runs. For comparison, the interconnectivity of the ideal optimal geometric sampling is also presented (*bottom*). The vertical dashed line marks the true period of the target.

scheduled, then it is very likely that similar further observations are also likely to be scheduled, if and when they become eligible according to the request constraints. This led to six close pairs of observations for which the second observation adds very little phase information to the lightcurve. These are therefore effectively wasted. However, the fact that the submission window was actually much wider than the positions chosen indicates that the telescope would not have made these observations under other circumstances (if the windows had been smaller, for example).

This behaviour is quite reasonable for a scheduler concerned with throughput, but illustrates the difficult balancing act demanded of an astronomer wishing to achieve cadenced observations. If the astronomer's constraints are too strict, the sequence is unlikely to be successful (there are few windows of opportunity). The more relaxed the constraints are made, the more opportunities there are for the scheduler to place the observations, but conversely, the more likely it is that the observations will not be placed with an appropriate spacing.

Ultimately, the true measure of data quality must be determined by frequency analysis. Periodograms for the adaptive agent run and the control run are presented as Figures 4.22 and 4.23 respectively.

The first thing to note is that based on the highest peak in the periodogram, both the ASA and the control runs correctly recover the true period of the signal. The differences lie in the relative strength of the true peak with respect to the rest of the periodogram. Calculating the peak ratio A gives $A = 1.40$ for the ASA dataset, and $A = 1.21$ for the control run. This makes it clear that the identification of the signal in the ASA dataset is more certain than in the control dataset. This is a relatively crude metric in the sense that it ignores the power present in the rest of the periodogram. The real win for the adaptive scheduling scheme is the conspicuous dampening of spurious spectral leakage across the full range of frequencies considered in the periodogram. This means that power in those other periods can be discounted with much more confidence than in the equivalent control run periodogram.

The periodogram, while powerful and convenient, nevertheless does not hold all the information required to differentiate between alternative periods. The periodogram is made up of a superposition of sinusoidal components, but the period of the observed target is not in general a pure sinusoid. Power in the periodogram can sometimes be attributed to possible sinusoidal fits to the dataset that are in fact quite unrealistic, and this can be determined by visual inspection of the lightcurve folded at one of these alternative periods.

Figures 4.24 and 4.25 illustrate the lightcurves that arise from folding on the other high peaks in the periodograms for the two runs. In the case of the ASA run (Fig. 4.24), the phase-folding on the second strongest peak after the true peak presents an unconvincing lightcurve. The third-strongest peak is harder to dismiss, but careful visual inspection would likely favour the true period, based on the overlap between error bars at equivalent phase positions. On the other hand, the second-highest peak in the periodogram of the control run cannot be easily ruled out using this criteria (Fig. 4.25, upper panel).

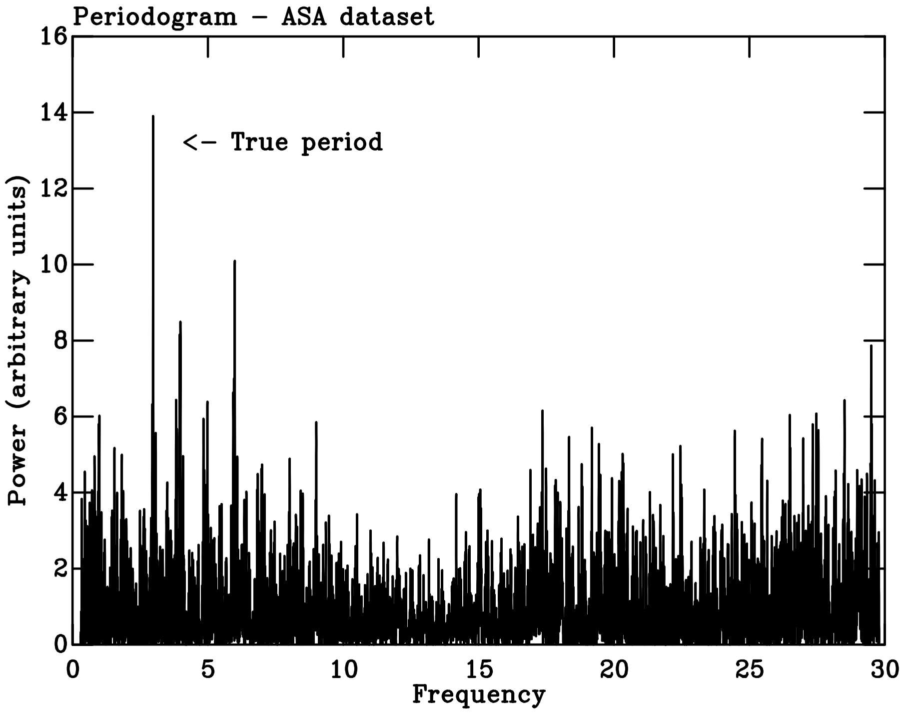

Figure 4.22: Periodogram of the ASA dataset. The highest peak corresponds to a period of 0.335898 days, in good agreement with the literature value. The value of the peak ratio is $A = 1.40$. Power at other frequencies is much lower than in the corresponding periodogram for the control run (Fig. 4.23, below).

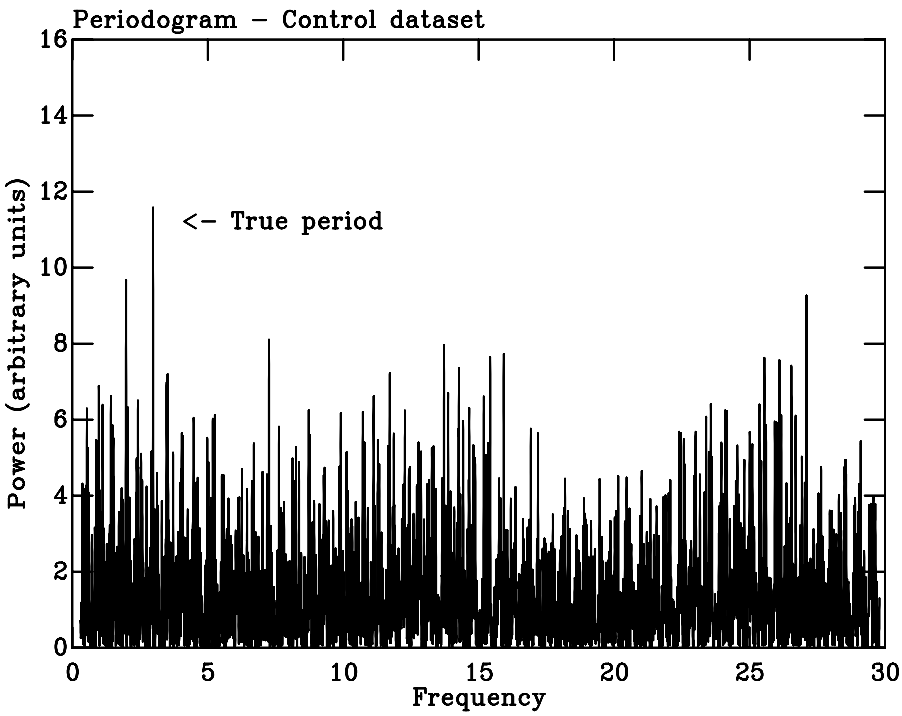

Figure 4.23: Periodogram of the control dataset. The highest peak corresponds to a period of 0.335670, in good agreement with the literature value. The value of the peak ratio is $A = 1.21$, somewhat poorer than for the ASA periodogram. In general, power at frequencies other than the true signal frequency is substantially greater than for the ASA run (Fig. 4.22).

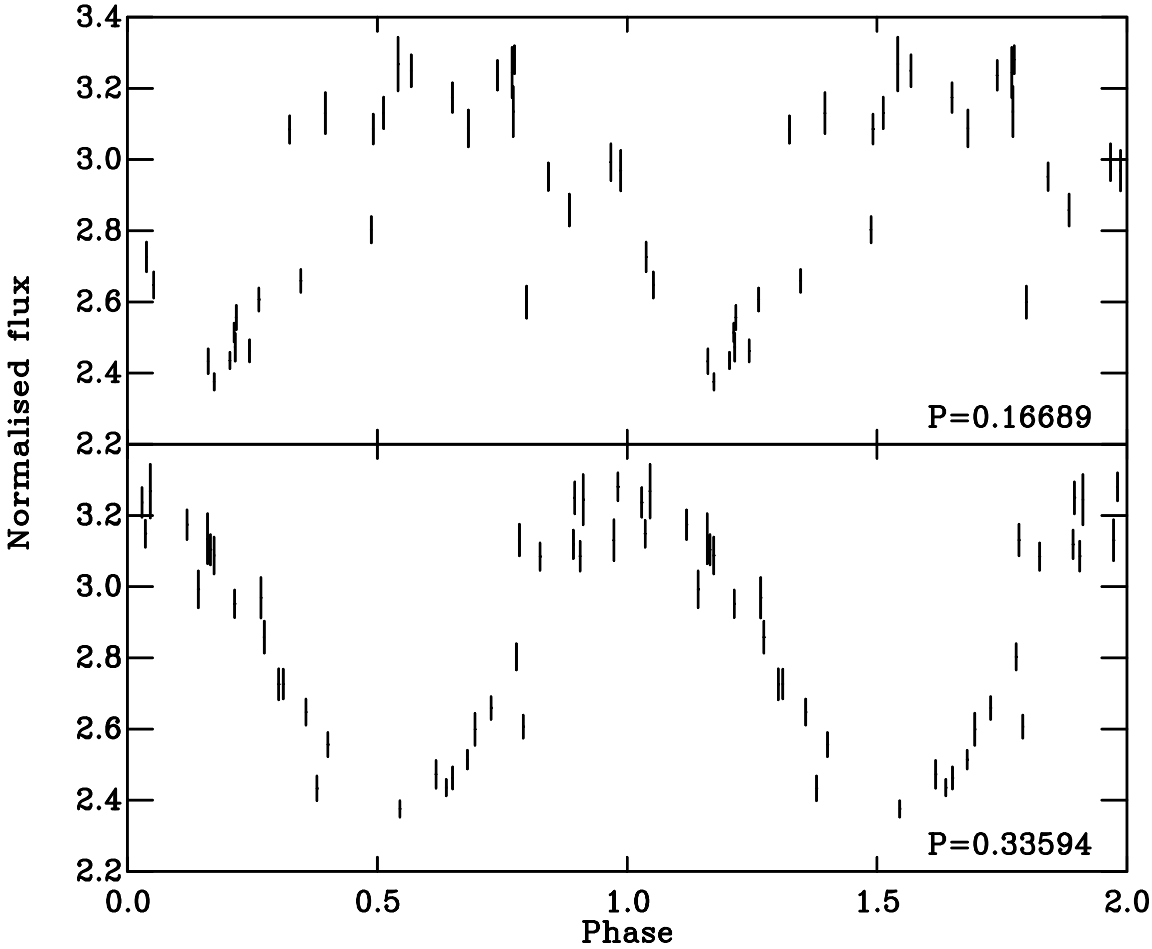

Figure 4.24: Folding the ASA dataset on the next highest peaks in the periodogram. The second highest peak (*upper panel*) occurs at a frequency of 5.9918/day (corresponding to a period of 0.16689 days). The third highest peak (*lower panel*) occurs at a frequency of 2.9767/day, corresponding to a period of 0.33594 days. In the first case, although considerable power is present in the periodogram, visual inspection of the folded lightcurve indicates that this potential period is incorrect. The third strongest frequency is not so easy to dismiss, because it is relatively close to the true period. One way is to consider the overlap in error bars between observations at similar phases. Careful comparison between these lightcurves is necessary to choose between them, but the strength of the true frequency in the periodogram is compelling.

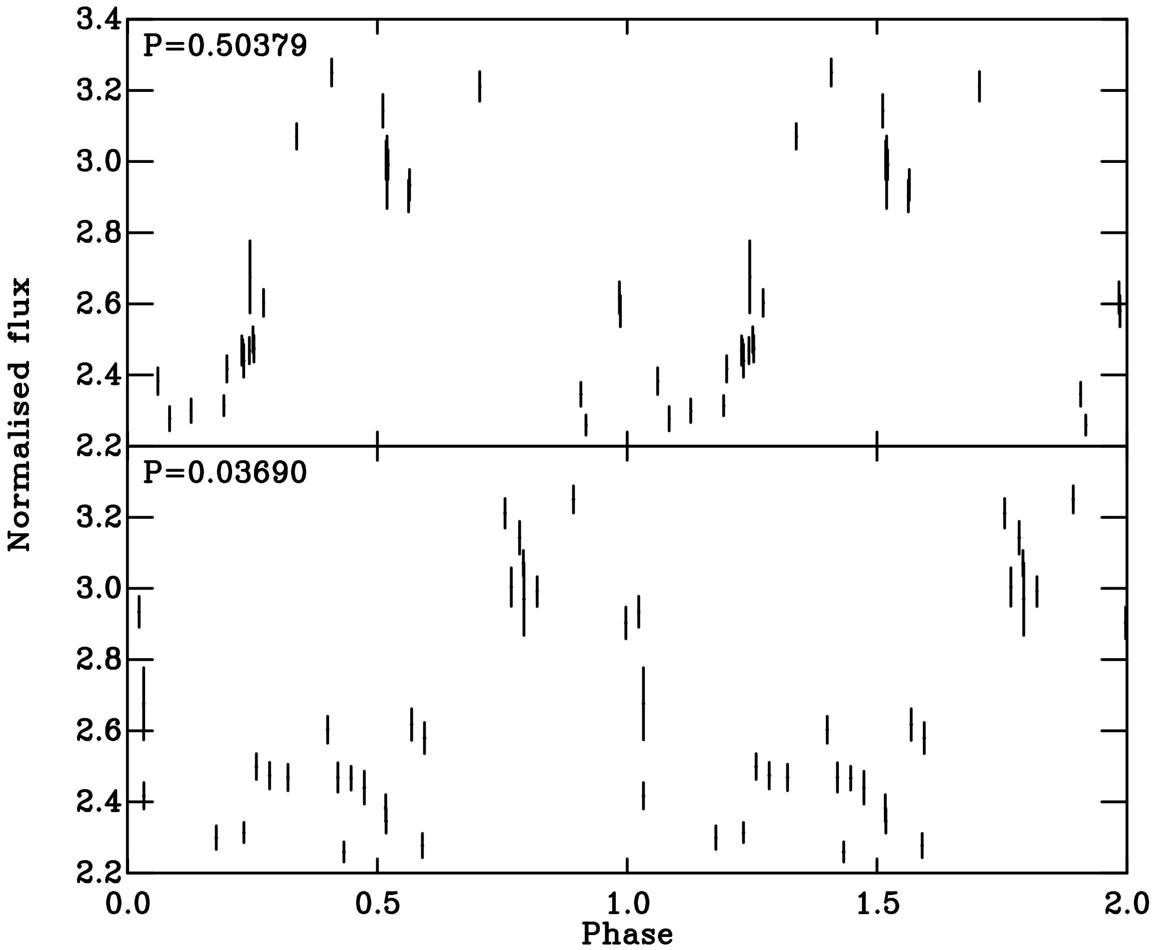

Figure 4.25: Folding the control dataset on the next highest peaks in the periodogram. The second highest peak (*upper panel*) occurs at a frequency of 1.9850/day (corresponding to a period of 0.50379 days). Although messy, a potential signal at this period cannot be ruled out by visual inspection. The third highest peak (*lower panel*) occurs at a frequency of 27.1007/day, corresponding to a period of 0.03690 days. In this case visual inspection of the folded lightcurve indicates that this potential period is incorrect.

4.6 Conclusions

In this chapter, an adaptive scheduling agent has been described that implements the optimal geometric sampling strategy for period searching described in Chapter 3. The agent performs its role in the context of an HTN-compliant network of telescopes (the eSTAR/Robonet-1.0 network), and is thus a proof-of-concept for this model of remote networked observing. The algorithm at the core of the agent automatically penalises the agent for choosing too many short gaps, but is similarly hostile to the chasing of large gaps when there are shorter, more valuable alternatives. In this way a balance point is determined that unambiguously defines the best option for the agent at any point in the run. The optimality calculation implicitly utilises the reordering property of optimal sequences, maximises the agent's cautiousness in the face of uncertainty, and is computationally cheap. It is also stable in the sense that any subset of acquired points are optimal, subject to the extant observing conditions. In the case of perfect observing behaviour (no failures), the optimal sequence is naturally recovered. The algorithm can respond to dynamic changes of the network of an unknown nature, because it makes no attempt to model the telescopes on which it relies.

The adaptive scheduling agent was used to run an observing programme on Robonet-1.0. A control run was carried out concurrently, based on an observing strategy of staggered observations typical of an astronomer using the Robonet-1.0 system. It was found that substantially more observations were acquired by the agent-based observing programme. More importantly, a large fraction of the gaps between temporally adjacent observations made by the agent observing programme were of the size specified by the optimal geometric sampling technique described in Chapter 3.

The data were reduced and lightcurves obtained for the two runs. The lightcurves were phase-folded on the known true period and compared. The agent-directed run had better phase coverage, both by eye and as quantified with the S metric. Fourier analysis of the two lightcurves demonstrated superior identification of the true period with respect to power elsewhere in the periodogram. The amplitude suppression of spurious peaks and broad-frequency spectral leakage in the agent-directed run indicates the successful minimisation of the window function predicted by the optimal geometric sampling theory. The contrast in periodogram when compared to the control run is marked.

The general conclusion of this chapter is that an adaptive scheduling agent that dynamically implements an optimal geometric sampling strategy works. The use of such an agent leads to substantial improvement in the quality of the dataset that may be obtained by an astronomer seeking to utilise a robotic telescope network for variability work.

Chapter 5

Conclusions

This thesis presents a practical solution to the problem of effective observation of time-varying phenomena using a network of robotic telescopes. Chapter 1 described the context of the work, charting the evolution of robotic observing, from its original inception as a handful of standalone instruments for automatic observing, to the present state-of-the-art, the interlinking of multiple telescopes and networks of telescopes to build an observing system greater than the sum of its parts.

Chapter 2 investigated temporal stellar variability in the young cluster h Per. The aims of this chapter were to explore a detailed example of the kind of time-domain observing this work addresses, showing the technical astronomy involved in observing and data reduction, as well as to demonstrate an example of the astrophysical insight that such a dataset can provide. Specifically, variability was used as a tracer for cluster membership, and it was argued that the sharp decrease in observed variables at higher parts of the sequence lends support to the hypothesis that the transition from a convective to a radiative stellar core leads to a change in the topology of the magnetic field at the surface of the star. The lack of variables at magnitudes brighter than the radiative-convective gap corresponding to this transition in the colour-magnitude diagram was presented as a potential tracer for this transition region. Finally, problems in the quality of the dataset arising from lack of phase coverage were identified as an example of a problem that could be successfully addressed by a robotic network.

Chapter 3 considered the theoretical problem of observation placement in the undersampled dataset regime, the common scenario for time-domain observations made using networked robotic telescopes. A set of metrics was developed to allow interesting properties of the sampling to be examined. Specifically, ways to quantify the phase coverage, the interconnectivity (a measure of the number of signal cycles sampled), and the strength of a given periodic signal relative to the strongest other signal in a periodogram (called the peak ratio) were presented. Simulations were performed to empirically investigate the properties of a number of sampling schemes. Analysis of the window functions generated from each sampling, and the application of these metrics to periodograms created from artificial datasets demonstrated that a geometric sampling scheme conclusively outperformed other sampling choices. To be effective the geometric scheme requires tailoring to the period range of interest, and a simple function to generate such a set of temporal positions was presented. It was also shown that the geometric series was largely insensitive to reordering, as long as the distribution of spacings was preserved, an important feature for practical application.

Chapter 4 applied the optimal sampling strategy to observe variability using a professional robotic telescope network. An autonomous software agent was designed and implemented that took the optimal sampling as the basis for an adaptive algorithm that could manage the uncertainties of a true observing environment dynamically, and without explicit user control. The reordering property of the optimal series provided the agent with crucial flexibility, and the ability to recalculate which reordered optimal sequence to pursue on-the-fly as a function of observation success or failure was fully utilised in the implementation. The observed target was the variable star BI Vir, and the data obtained were used to identify the period and evaluate the performance of the agent. It was found that the agent achieved a large fraction of its observations at the spacings dictated by the optimal sampling. Comparison with a simultaneous control run, placed with observing constraints typical of an astronomer using the network, demonstrated the significant performance gain provided by the agent. More observations were completed successfully and the spacing between them was much more effective, a fact demonstrated by comparing the phase coverage of the two runs as well as the quality of the final periodograms.

5.1 Further work

There are many directions in which this work may be taken. The optimal sampling strategy described in this thesis has been shown to be effective, and a possible phenomenological explanation in terms of sampled frequencies has been proposed. However, no attempt has been made to characterise the smoothness metric of the window function, or to analytically solve this to find the global minima. One such approach would be to seek a minimisation of the area under the window function beyond the main peak. After integrating, it may be possible to solve the resulting function in the limit as the area tends to zero.

Another approach to the sampling problem would be to exhaustively calculate the smoothness of the window function at every point of a discretised four observation simulation. For the dynamic range of frequencies assumed in the text, this should be a feasible calculation. The result would be a visualisation of the surface of the window function smoothness in two dimensions, and in particular, the variety and location of minima should be clearly defined. The shape of the surface may provide clues to the likely topology in higher dimensions, with the tantalising possibility of generalisation to sampling problems of arbitrary size.

A third way of tackling the search for minima would be to cast it in terms of a classical optimisation problem, and seek to numerically locate the position of global minima through a limited number of trials which are placed to take maximum advantage of the assumed correlation between similar points in the phase space. Work is ongoing in this area. Algorithms that are currently being investigated include Nelder-Mead simplex (Nelder & Mead, 1964) and particle swarm optimisation (Eberhart & Kennedy, 1995).

An alternative approach, currently under investigation by the author, is to analytically identify the correct frequency distribution required for optimal undersampling, and then to determine the correct distribution of points through general arguments relating Fourier space to the time domain. For example, minimising structure in the window function may, in the limit, be equivalent to choosing a delta function at $\nu = 0$, and setting all other frequencies to zero. While this condition may not be achievable in practice, the limit may provide insight as to the 'maximum unevenness' that a set of points can provide in practice.

Work on similar problems exists in the field of interferometry. An interferometer is made up of a number of spatially separated telescopes which, through the technique of aperture synthesis, can have their signals combined. The resulting network has the same angular resolution as a single instrument with a diameter equal to the longest baseline between telescopes. The pattern of gaps between receivers determines the sampling of the *visibility*, the set of amplitude and phase measurements observed by the interferometer. The Fourier transform of this complex quantity is the brightness distribution of the object, i.e. the image. Thus each measurement of the visibility can be considered a measurement of one of the Fourier components of the resulting image. The position of each receiver therefore determines which components are sampled. Significant work has been performed to determine the ideal positions for the individual receivers, and the resulting configurations are called *minimum redundancy linear arrays* (Moffet, 1968; Ishiguro, 1980). They are arranged to sample every wavelength up to some maximum value. Some wavelengths are unhelpfully sampled more than once; the array spacing aims to minimise these repeated measurements. Comparison of this problem with the time domain problem is in progress by the author.

An obvious step would be to try and improve the reliability of the scoring mechanism. One way to do this would be for an agent to track the history of past requests to individual telescopes, and compare the score returned with the actual outcome of the observation request. This would effectively provide a way to 'normalise' the scores provided by each telescope according to a trust factor. Since this trust factor is calculated heuristically, this approach is not a function of the scoring algorithm and requires no explicit knowledge of the telescope scheduler's decision-making process. Additionally it is dynamic, reflecting changes in scoring behaviour promulgated over time. Variants of such a learning mechanism could also be applied to other agent experiences, such as the effective availability of different telescopes at different times of the year.

Another direction for further work is applying currently unused portions of the HTN standard. Although not used by this agent, a 'differential' score, which provides a continuum of observation likelihoods over a range of time, is returned by the Robonet-1.0 telescopes. This extra information could potentially allow higher success rates by increasing the potential for fine tuning by the agent.

Extending the HTN protocol is another way to gain flexibility. For example, if an agent were able to check on the score of a queued observation, alternatives to the 'submit and see' strategy become viable. For example, an agent could submit observations to many telescopes, leave them queued, and, when the time for the observation to begin is near, send abort messages to all telescopes other than the highest score (T. Naylor, priv. comm.). From the agent perspective, this is functionally equivalent to waiting until the last moment to place an observation, a technique which is advantageous to the agent but harmful to a telescope scheduler. One way to discourage such tactics, with a view to increasing the reliability of individual schedulers, would be to penalise agents that submit observation requests close to the start time of the observation (Naylor, 2008).

The experiment of running the adaptive scheduling agent showed that against the current Liverpool scheduler, a reactive mode may achieve a higher observation success rate for an individual observer by allowing finer control over observation submission. This becomes particularly pertinent for the problem of long time series spread across several HTN telescopes. In this case, there is no overarching scheduler responsible for managing the continuity of the time series, and therefore the burden of achieving observations with the correct cadence falls on the user. This would be an ideal case for a reactive agent to handle. Examples include long-term monitoring of variability, and monitoring of potential microlensing candidates with the aim of identifying the anomolous outburst that would signify the prescence of an extra-solar planet around a distant star.

The reactive component of the adaptive scheduling agent is required because of the uncertainty associated with achieving any requested observation. This is a specific example of the general problem of acheiving a fully specified, arbitrarily-spaced set of time series observations. If a telescope scheduling system supports regular cadenced time series requests, which are a common way of specifying monitoring programmes, then it is only a small step to support arbitrary time series. Internally, the scheduler must handle the problem of observation uncertainty, but this is not specific to the undersampling problem solved by the adaptive scheduling agent. In such a situation, therefore, the user would no longer require a reactive component in order to schedule undersampled observation sequences — the technical details of actually acquiring the observations would be firmly in the domain of the telescope scheduler. This would remove much of the risk and uncertainty associated with the current mode of observing, as well as drastically simplifying the task of the user.

Bibliography

Allan A., Bischoff K., Burgdorf M., Cavanagh B., Christian D., Clay N., Dickens R., Economou F., Fadavi M., Frazer S., Granzer T., Grosvenor S., Hessman F. V., Jenness T., Koratkar A., et al. 2006, in Observatory Operations: Strategies, Processes, and Systems. Edited by Silva, David R.; Doxsey, Rodger E.. Proceedings of the SPIE, Volume 6270, pp. 62700H (2006). Operating a heterogeneous telescope network

Allan A., Hessman F., Bischoff K., Burgdorf M., Cavanagh B., Christian D., Clay N., Dickens R., Economou F., Fadavi M., Fraser S., Granzer T., Grosvenor S., Jenness T., Koratkar A., Lehner M., Mottram C., et al. 2006, Astronomische Nachrichten, 327, 744

Allan A., Naylor N., Steele I., Carter D., Etherton J., Mottram C., 2003, in Payne H. E., Jedrzejewski R. I., Hook R. N., eds, Astronomical Data Analysis Software and Systems XII Vol. 295 of Astronomical Society of the Pacific Conference Series, eSTAR: Building an Observational GRID. pp 13–+

Allan A., Naylor T., Saunders E. S., 2006, Astronomische Nachrichten, 327, 767

Allan A., Naylor T., Steele I., Carter D., Jenness T., Economou F., Adamson A., 2004b, in Ochsenbein F., Allen M. G., Egret D., eds, ASP Conf. Ser. 314: Astronomical Data Analysis Software and Systems (ADASS) XIII eSTAR: Astronomers, Agents and when Robotic Telescopes aren't.... pp 597–+

Allan A., Naylor T., Steele I. A., Jenness T., Cavanagh B., Economou F., Saunders E., Adamson A., Etherton J., Mottram C., 2004a, in Lewis H., Raffi G., eds, Advanced Software, Control, and Communication Systems for Astronomy. Edited by Lewis, Hilton; Raffi, Gianni. Proceedings of the SPIE, Volume 5496, pp. 313-322 (2004). eSTAR: intelligent observing and rapid responses. pp 313–322

Allan et al. A., 2008, Astronomische Nachrichten (in prep.)

Antonioli P., Fienberg R. T., Fleurot F., Fukada Y., Fulgione W., Habig A., Heise J., McDonald A. B., Mills C., Namba T., Robinson L. J., Scholberg K., Schwendener M., Sinnott R. W., Stacey B., Suzuki Y., et al. 2004, New J. Phys., p. 114

Ashley M. C. B., Burton M. G., Lawrence J. S., Storey J. W. V., 2004, Astronomische Nachrichten, 325, 619

Barthelmy S. D., Cline T. L., Butterworth P., Kippen R. M., Briggs M. S., Connaughton V., Pendleton G. N., 2000, in Kippen R. M., Mallozzi R. S., Fishman G. J., eds, AIP Conf. Proc. 526: Gamma-ray Bursts, 5th Huntsville Symposium GRB Coordinates Network (GCN): A Status Report. pp 731–+

Barthelmy S. D., Cline T. L., Gehrels N., Bialas T. G., Robbins M. A., Kuyper J. R., Fishman G. J., Kouveliotou C., Meegan C. A., 1994, in Fishman G. J., ed., AIP Conf. Proc. 307: Gamma-Ray Bursts BACODINE: The Real-Time BATSE Gamma-Ray Burst Coordinates Distribution Network. pp 643–+

Beckwith S. V. W., Sargent A. I., Chini R. S., Guesten R., 1990, AJ, 99, 924

Bertout C., Basri G., Bouvier J., 1988, ApJ, 330, 350

Beutler F. J., 1970, IEEE Trans. Information Theory, 16, 147

Bode M. F., Cardiff U. U. Hertfordshire U. Leicester St Andrews U. Queens U. B., Mullard Space Science Lab. U. Exeter U. Southampton U. Manchester 2004, in Bulletin of the American Astronomical Society RoboNet-1.0: A Prototype Global Network of Large Robotic Telescopes. pp 1400–+

Boyd L. J., Epand D., Bresina J., Drummond M., Swanson K., Crawford D. L., Genet D. R., Genet R. M., Henry G. W., McCook G. P., Neely W., Schmidtke P., Smith D. P., Trueblood M., 1993, International Amateur-Professional Photoelectric Photometry Communications, 52, 23

Bragg A. E., Kenyon S. J., 2005, AJ, 130, 134

Bratko I., 1990, Prolog programming for artificial intelligence, 2nd edn. Addison Wesley

Bray T., Paoli J., Sperberg-McQueen C. M., Maler E., Yergeau F., Cowan J., , 1998, Extensible Markup Language (XML) 1.1 (Second Edition) at `www.w3.org/TR/xml11`

Brooks R. A., 1991, in Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI-91) Intelligence without reason. Sydney, Australia, pp 569–595

Burrows D. N., Hill J. E., Nousek J. A., Kennea J. A., Wells A., Osborne J. P., Abbey A. F., Beardmore A., Mukerjee K., Short A. D. T., Chincarini G., Campana S., Citterio O., Moretti A., et al. 2005, Space Science Reviews, 120, 165

Chaplin W. J., Elsworth Y., Howe R., Isaak G. R., McLeod C. P., Miller B. A., van der Raay H. B., Wheeler S. J., New R., 1996, Sol. Phys., 168, 1

Chavez A., Maes P., 1996, in Proceedings of the 1st International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM-96) Kasbah: an agent marketplace for the buying and selling of goods. London, UK, pp 75–90

Christian D. J., Pollacco D. L., Skillen I., Irwin J., Clarkson W. I., Collier Cameron A., Evans N., Fitzsimmons A., Haswell C. A., Hellier C., Hodgkin S. T., Horne K., Kane S. R., Keenan F. P., Lister T. A., Norton A. J., Osborne J., et al. 2006, Astronomische Nachrichten, 327, 800

Code A. D., 1992, in Filippenko A. V., ed., ASP Conf. Ser. 34: Robotic Telescopes in the 1990s The Wisconsin APT - The first robotic telescope. pp 3–8

Colgate S. A., Moore E. P., Carlson R., 1975, PASP, 87, 565

Collier Cameron A., 2001, in Boffin H. M. J., Steeghs D., Cuypers J., eds, Astrotomography, Indirect Imaging Methods in Observational Astronomy Vol. 573 of Lecture Notes in Physics, Berlin Springer Verlag, Spot Mapping in Cool Stars. pp 183–+

Comerón F., Mathys G., Kaufer A., Hainaut O., Hanuschik R., Romaniello M., Silva D., Quinn P., 2006, in Observatory Operations: Strategies, Processes, and Systems. Edited by Silva, David R.; Doxsey, Rodger E.. Proceedings of the SPIE, Volume 6270, pp. 62700X (2006). VLT service mode operations at seven years

Crawford D. L., 1988, in Bulletin of the American Astronomical Society Advantages of a Global Network of Automatic Small Telescopes. pp 675–+

Crawford D. L., 1992, in Adelman S. J., Dukes Jr. R. J., Adelman C. J., eds, ASP Conf. Ser. 28: Automated Telescopes for Photometry and Imaging GNAT: Global Network of Automated Telescopes. pp 123–+

Crawford D. L., Glaspey J. W., Perry C. L., 1970, AJ, 75, 822

Darley V., 1994, in Brooks R., Maes P., eds, Artificial Life IV, Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems Emergent phenomena and complexity. MIT Press, pp 411–416

Deeming T. J., 1975, Ap&SS, 36, 137

Diego F., 1985, PASP, 97, 1209

Drummond M., Bresina J., Edgington W., Swanson K., Henry G., Drashcer E., 1995, in Henry G. W., Eaton J. A., eds, ASP Conf. Ser. 79: Robotic Telescopes. Current Capabilities, Present Developments, and Future Prospects for Automated Astronomy Flexible Scheduling of Automatic Telescopes Over the Internet. pp 101–+

Dworetsky M. M., 1983, MNRAS, 203, 917

Eberhart R., Kennedy J., 1995, in Proceedings of the Sixth International Symposium on Micro Machine and Human Science "a new optimizer using particle swarm theory". IEEE Service Center, Piscataway, NJ, Nagoya, Japan, pp 39–43

Economou F., Jenness T., Cavanagh B., Adamson A. J., Allan A., 2006, Astronomische Nachrichten, 327, 788

Erera A., Morales J., Savelsbergh M. W. P., 2005, Transportation Research Part E: The Logistics and Transportation Review, 41, 551

ESO 2004, in Minutes of the 44th STC meeting VLT/VLTI Science Operations Policy

Etherton J., Steele I. A., Mottram C. J., 2004, in Quinn P. J., Bridger A., eds, Optimizing Scientific Return for Astronomy through Information Technologies. Edited by Quinn, Peter J.; Bridger, Alan. Proceedings of the SPIE, Volume 5493, pp. 90-96 (2004). A free market in telescope time?. pp 90–96

Evans D. W., 2001, Astronomische Nachrichten, 322, 347

Eyer L., Bartholdi P., 1999, A&AS, 135, 1

Foster I., Kesselman C., eds, 1999, The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, p. 330

Fraser S., Steele I. A., 2004, in Quinn P. J., Bridger A., eds, Optimizing Scientific Return for Astronomy through Information Technologies. Edited by Quinn, Peter J.; Bridger, Alan. Proceedings of the SPIE, Volume 5493, pp. 331-340 (2004). Robotic telescope scheduling: the Liverpool Telescope experience. pp 331–340

Fraser S. N., 2006, Astronomische Nachrichten, 327, 779

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748

Gladwell M., 1998, The Tipping Point: How Little Things Can Make a Big Difference. Back Bay Books

Gotz D., Mereghetti S., Shaw S., Beck M., Borkowski J., 2004, GRB Coordinates Network, 2866, 1

Granzer T., 2004, Astronomische Nachrichten, 325, 513

Granzer T., 2006, Astronomische Nachrichten, 327, 792

Guth A. H., 1981, Physical Review D, 23

Hartmann L., 2001, Accretion Processes in Star Formation. Accretion Processes in Star Formation, by Lee Hartmann, pp. 237. ISBN 0521785200. Cambridge, UK: Cambridge University Press, January 2001.

Hayashi C., 1966, ARA&A, 4, 171

Helmer L., Morrison L. V., 1985, Vistas in Astronomy, 28, 505

Henden A. A., Stone R. C., 1998, AJ, 115, 296

Herbst W., Maley J. A., Williams E. C., 2000, AJ, 120, 349

Hessman F. V., 2006a, Astronomische Nachrichten, 327, 763

Hessman F. V., 2006b, Astronomische Nachrichten, 327, 751

Hill G. J., 2000, in Iye M., Moorwood A. F., eds, Proc. SPIE Vol. 4008, p. 50-61, Optical and IR Telescope Instrumentation and Detectors, Masanori Iye; Alan F. Moorwood; Eds. Hobby-Eberly Telescope: instrumentation and current performance. pp 50–61

Horne J. H., Baliunas S. L., 1986, ApJ, 302, 757

Horne K., 2003, in Deming D., Seager S., eds, ASP Conf. Ser. 294: Scientific Frontiers in Research on Extrasolar Planets Status aand Prospects of Planetary Transit Searches: Hot Jupiters Galore. pp 361–370

Horne K., Wade R. A., Szkody P., 1986, MNRAS, 219, 791

Huhns M., Singh M. P., eds, 1998, Readings in Agents. Morgan Kaufmann

Humphries C. M., Purkins T. E., 1980, Endeavour, 4, 132

Ishiguro M., 1980, Radio Science, 15, 1163

Jeans J. H., 1902, Phil. Trans. R. Soc., 199

Jennings N. R., Corera J. M., Laresgoiti I., Mamdani E. H., Perriollat F., Skarek P., Varga L. Z., 1996, IEEE Expert, 11, 60

Kant I., 1755, Allgemeine Naturgeschichte und Theorie des Himmels (Universal Natural History and Theory of Heaven). Reprinted in Kant's gesammelte Schriften (Kant's Collected Works), Berlin: G. Reimer, 1905

Kenyon S. J., Hartmann L., 1987, ApJ, 323, 714

Kipling R., 1902, "*I keep six honest serving men...*", Just So Stories. Macmillan, London

Koen C., 2006, MNRAS, 371, 1390

Köenigl A., 1991, ApJL, 370, L39

Kurtz D. W., 1983, Informational Bulletin on Variable Stars, 2285, 1

Laplace P.-S., 1796, Exposition du Système du Monde (The System of the World). Paris : V. Courcier

Larson R. B., 2003, Reports of Progress in Physics, 66, 1651

Lehner M. J., Alcock C., Axelrod T., Bianco F., Byun Y.-I., Chen W.-P., Cook K. H., Dave R., de Pater I., Giammarco J., King S.-K., Lee T., Lissauer J., Marshall S. L., Mondal S., Nihei T., Rice J., Schwamb M., et al. 2006, Astronomische Nachrichten, 327, 814

Littlefair S., Naylor T., Burningham B., Jeffries R. D., 2005, MNRAS, 358, 341

Ljunberg M., Lucas A., 1992, in Proceedings of the 2nd Pacific Rim International Conference on AI (PRICAI-92) The OASIS air traffic management system. Seoul, Korea

Lomb N. R., 1976, Ap&SS, 39, 447

Lynden-Bell D., Pringle J. E., 1974, MNRAS, 168, 603

Maran S. P., 1967, Science, 158, 867

Mayne N. J., Naylor T., Littlefair S. P., Saunders E. S., Jeffries R. D., 2007, MNRAS, 375, 1220

McNall J. F., Miedaner T. L., Code A. D., 1968, AJ, 73, 756

Moffet A. T., 1968, IEEE Trans. Ant. Prop., 16, 172

Mould J. R., Hughes S. M. G., Stetson P. B., Gibson B. K. e. a., 2000, ApJ, 528, 655

Nather R. E., Winget D. E., Clemens J. C., Hansen C. J., Hine B. P., 1990, ApJ, 361, 309

Naylor T., 1998, MNRAS, 296, 339

Naylor T., 2008, Astronomische Nachrichten (in prep.)

Naylor T., Allan A., Steele I. A., 2006, Astronomische Nachrichten, 327, 741

Naylor T., Bath G. T., Charles P. A., Hassal B. J. M., Sonneborn G., van der Woerd H., van Paradijs J., 1988, MNRAS, 231, 237

Naylor T., Totten E. J., Jeffries R. D., Pozzo M., Devey C. R., Thompson S. A., 2002, MNRAS, 335, 291

Nelder J., Mead R., 1964, The Computer Journal, 7, 308

Newton I., 1756, Four Letters from Sir Isaac Newton to Doctor Bentley containing some Arguments in Proof of a Deity (Letter I). Reprinted in I. B. Choen (ed.), Isaac Newton's Papers & Letters on Natural Philosophy, 2nd edition. Cambridge, Mass., and London, 1978., pp 279–312

Nyquist H., 1928, Trans. AIEE, 47, 617

Oosterhoff P. T., 1937, Annalen van de Sterrewacht te Leiden, 17, 1

Ormerod P., Campbell M., 1998, The Evolution of Family Structures in a Social Context. Volterra Consulting

Østergaard P. A., 2005, Renewable Energy, 30, 977

Pennypacker C., Boer M., Denny R., Hessman F. V., Aymon J., Duric N., Gordon S., Barnaby D., Spear G., Hoette V., 2002, A&A, 395, 727

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical recipes in FORTRAN. The art of scientific computing. Cambridge: University Press, —c1992, 2nd ed.

Quimby R., Höflich P., Kannappan S. J., Rykoff E., Rujopakarn W., Akerlof C. W., Gerardy C. L., Wheeler J. C., 2006, ApJ, 636, 400

Rees P. C. T., Conway P. B., Mansfield A. G., Mucke-Herzberg D., Rosing W., Surrey P. J., Taylor S., 2006, in Ground-based and Airborne Telescopes. Edited by Stepp, Larry M.. Proceedings of the SPIE, Volume 6267, pp. 62673S (2006). A global network of robotic telescopes

Roberts D. H., Lehar J., Dreher J. W., 1987, AJ, 93, 968

Rosenschein S. J., Kaelbling L. P., 1996, in P. E. Agre and S.J. Rosenschein ed., , Computational Theories of Interaction and Agency. MIT Press, pp 515–540

Saunders E. S., Naylor T., Allan A., 2006a, Astronomische Nachrichten, 327, 783

Saunders E. S., Naylor T., Allan A., 2006b, A&A, 455, 757

Saunders E. S., Naylor T., Mayne N., Littlefair S. P., 2008, in prep.

Scargle J. D., 1982, ApJ, 263, 835

Schelling T. C., 1978, Micromotives and Macrobehaviour. WW Norton, New York

Schuster A., 1898, Terrestrial Magnetism, 3, 13

Schwarzenberg-Czerny A., 2003, in ASP Conf. Ser. 292: Interplay of Periodic, Cyclic and Stochastic Variability in Selected Areas of the H-R Diagram An astronomer's guide to period searching. pp 383–+

Seaman R., Allan A., Barthelmy S., Drake A., Graham M., White R., Warner P., , 2006, VOEvent Transport Protocol Agreement v.1.1 (draft) at `http://www.ivoa.net/internal/IVOA/IvoaVOEvent/VOEvent_transport_060412.html`

Seaman R., Williams R., Allan A., Barthelmy S., Bloom J., Graham M., Hessman F., Marka S., Rots A., Stoughton C., Vestrand T., White R., Wozniak P., , 2005, Sky Event Reporting Metadata (VOEvent) at `www.ivoa.net/Documents/latest/VOEvent.html`

Shannon C. E., 1949, Proc. Inst. of Radio Engineers, 37, 10

Shu F. H., 1977, ApJ, 214, 488

Slesnick C. L., Hillenbrand L. A., Massey P., 2002, ApJ, 576, 880

Smith A., 1776, An Inquiry into the Nature and Causes of the Wealth of Nations

Smith A. M. S., Collier Cameron A., Christian D. J., Clarkson W. I., Enoch B., Evans A., Haswell C. A., Hellier C., Horne K., Irwin J., Kane S. R., Lister T. A., Norton A. J., Parley N., Pollacco D. L., Ryans R., Skillen I., Street R. A., et al. 2006, MNRAS, 373, 1151

Steele I. A., 2001, Astronomische Nachrichten, 322, 307

Steele I. A., Clay N. R., Cohen J., Lee W., Mottram C. J., 2006, in Observatory Operations: Strategies, Processes, and Systems. Edited by Silva, David R.; Doxsey, Rodger E.. Proceedings of the SPIE, Volume 6270, pp. 62700M (2006). Vol. 6270 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, A free market in telescope time II: a test implementation

Steele I. A., Smith R. J., Rees P. C., Baker I. P., Bates S. D., Bode M. F., Bowman M. K., Carter D., Etherton J., Ford M. J., Fraser S. N., Gomboc A., Lett R. D. J., Mansfield A. G., Marchant J. M., et al. 2004, in Oschmann Jr. J. M., ed., Ground-based Telescopes. Edited by Oschmann, Jacobus M., Jr. Proceedings of the SPIE, Volume 5489, pp. 679-692 (2004). The Liverpool Telescope: performance and first results. pp 679–692

Stellingwerf R. F., 1978, ApJ, 224, 953

Stobie R., Meiring J., Buckley D. A., 2000, in Dierickx P., ed., Proc. SPIE Vol. 4003, p. 355-362, Optical Design, Materials, Fabrication, and Maintenance, Philippe Dierickx; Ed. Design of the Southern African Large Telescope (SALT). pp 355–362

Stolte A., Brandner W., Brandl B., Zinnecker H. ., Grebel E. K., 2004, AJ, 128, 765

Strom K. M., Strom S. E., Edwards S., Cabrit S., Skrutskie M. F., 1989, AJ, 97, 1451

Swedenborg E., 1734, Opera Philosophica et Mineralia (Philosophical and Mineralogical Works), Principia, Volume I. Dresden and Leipzig: F. Hekel

Sweeney D. W., 2006, in Ground-based and Airborne Telescopes. Edited by Stepp, Larry M.. Proceedings of the SPIE, Volume 6267, pp. 626706 (2006). Overview of the Large Synoptic Survey Telescope project

Tapia M., Roth M., Costero R., Navarro S., 1984, Revista Mexicana de Astronomia y Astrofisica, 9, 65

Vestrand W. T., Borozdin K. N., Brumby S. P., Casperson D. E., Fenimore E. E., Galassi M. C., McGowan K., Perkins S. J., Priedhorsky W. C., Starr D., White R., Wozniak P., Wren J. A., 2002, in Kibrick R. I., ed., Advanced Global Communications Technologies for Astronomy II. Edited by Kibrick, Robert I. Proceedings of the SPIE, Volume 4845, pp. 126-136 (2002). The RAPTOR experiment: a system for monitoring the optical sky in real time. pp 126–136

Vestrand W. T., Theiler J., Woznia P. R., 2004, Astronomische Nachrichten, 325, 477

Vestrand W. T., Wozniak P. R., Wren J. A., Fenimore E. E., Sakamoto T., White R. R., Casperson D., Davis H., Evans S., Galassi M., McGowan K. E., Schier J. A., Asa J. W., Barthelmy S. D., Cummings J. R., Gehrels N., Hullinger D., et al. 2005, Nature, 435, 178

Vogt S. S., Penrod G. D., Hatzes A. P., 1987, ApJ, 321, 496

Waelkens C., Lampens P., Heynderickx D., Cuypers J., Degryse K., Poedts S., Polfliet R., Denoyelle J. an d van den Abeele K., Rufener F., Smeyers P., 1990, A&AS, 83, 11

Wall J. V., Jenkins C. R., 2003, Practical Statistics for Astronomers. Princeton Series in Astrophysics

Wall L., Christiansen T., Orwant J., 2000, Programming Perl. O'Reilly

Wallace P. T., 1994, in Crabtree D. R., Hanisch R. J., Barnes J., eds, Astronomical Data Analysis Software and Systems III Vol. 61 of Astronomical Society of the Pacific Conference Series, The SLALIB Library. pp 481–+

White R. R., Allan A., Barthelmy S., Bloom J., Graham M., Hessman F. V., Marka S., Rots A., Scholberg K., Seaman R., Stoughton C., Vestrand W. T., Williams R., Wozniak P. R., 2006, Astronomische Nachrichten, 327, 775

White R. R., Evans S. M., Vestrand W. T., Warren M. S., Wren J. A., Wozniak P. R., 2006, Astronomische Nachrichten, 327, 758

Wooldridge M., 2002, An Introduction to Multiagent Systems. Wiley

Wooldridge M., Jennings N. R., 1995, "The Knowledge Engineering Review", 10, 115

Woźniak P. R., Vestrand W. T., Akerlof C. W., Balsano R., Bloch J., Casperson D., Fletcher S., Gisler G., Kehoe R., Kinemuchi K., Lee B. C., Marshall S., McGowan K. E., McKay T. A., Rykoff E. S., Smith D. A., Szymanski J., Wren J., 2004, AJ, 127, 2436

Yost S. A., Aharonian F., Akerlof C. W., Ashley M. C. B., Barthelmy S., Gehrels N., Göğüş E., Güver T., Horns D., Kızıloğlu Ü., Krimm H. A., McKay T. A., Özel M., Phillips A., Quimby R. M., et al. 2006, Astronomische Nachrichten, 327, 803

Index