A Framework for Spatio-Temporal Data Analysis and Hypothesis Exploration*

Campbell, Alex[a], B. Pham[a] and Y.-C Tian[a]

[a]Computational Intelligence Group, Faculty of Information Technology
Queensland University of Technology, GPO Box 2434, Brisbane Qld 4001, Australia. E-Mail:
ab.campbell@qut.edu.au

Abstract: We present a general framework for pattern discovery and hypothesis exploration in spatio-temporal data sets that is based on delay-embedding. This is a remarkable method of nonlinear time-series analysis that allows the full phase-space behaviour of a system to be reconstructed from only a single observable (accessible variable). Recent extensions to the theory that focus on a probabilistic interpretation extend its scope and allow practical application to noisy, uncertain and high-dimensional systems. Our framework uses these extensions to aid alignment of spatio-temporal sub-models (hypotheses) to empirical data - for example, satellite images plus remote-sensing - and to explore behaviours consistent with this alignment. The novel aspect of the work is a mechanism for linking global and local dynamics using a holistic spatio-temporal feedback loop. An example framework is devised for an urban planning application, transit-oriented developments, and its feasibility is demonstrated with real data.

Keywords: Spatio-temporal, data mining, hypothesis exploration, delay-embedding.

1 INTRODUCTION

The continual acceleration of satellite imaging and evolution of remote-sensing techniques provide massive amounts of data, and potentially massive amounts of knowledge. Because space-time data is based in the same space-time we are so immersed in, it is perhaps more conceptually intuitive than that contained in standard data warehouses. At the same time, traditional data analysis techniques that ignore the spatio-temporal location of data attributes perform poorly. Thus spatio-temporal data analysis is both subtly and significantly different. As a research field it is young and, though rapidly developing, much of the territory is unmapped.

Most current approaches to spatio-temporal data analysis focus on a two-step process: identifying objects in space, then tracking them in time. Good examples of these are the combination of a Kalman filter (for tracking) with spatial analysis techniques such as Kriging (Huang and Cressie [1996]), or Markov Random Fields (Chawla [2003]). This is an intuitively reasonable approach, however there are already some fundamental restrictions imposed by such a reductionist treatment of space-time. Consider a bush-fire scenario where, until the fire has run out of flammable material or been extinguished, there will be a front moving over the landscape. Imagine that a satellite is monitoring the fire and taking images: for each instantaneous image, the corresponding pattern would be a step function at the location of the front. Even if the proper static patterns can be identified, a dynamical description would have to incorporate a vast number of them, one that is exponential in the size of the system. This is because the front may be moving over the whole extent of the system - re-use of region-limited patterns is not an option. [1]

Another of the more significant obstacles to a systematic analysis of space-time data is that of scaling. For example results obtained at one particular level of detail (resolution) may be no longer valid at different levels. The incorporation of domain knowledge in particular is hampered by inadequate consideration of scale-dependence, because it often is patchy at any particular resolution and yet is poten-

*This work is supported by an Australian Research Council Linkage grant in collaboration with the Built Environment Research Unit (BERU), Queensland Department of Public Works. We gratefully acknowledge BERU and the Planning and Information Forecasting Unit of the Department of Local Government and Planning for contribution of data sets.

[1]This scenario is modified from a more abstract example in Kantz and Schrieber [1997], p. 247

tial very powerful when considered across all scales.

We present a general framework for pattern discovery in spatio-temporal data sets that extends out from time-series analysis techniques, and focuses on non-autonomous dynamics. The novel aspect of the work is a mechanism for linking global and local dynamics using a holistic spatio-temporal feedback loop. The holistic space-time approach avoids the problem of explosion in number of patterns to track, and the feedback between local and global scales facilitates integration of domain knowledge.

We consider an urban planning application to illustrate the framework. This is a good test domain because it is highly active in both space and time, it requires global understanding yet is extremely data rich on local scales, and there is both scope and need for integration with domain-knowledge. This paper builds on work presented in Campbell et al. [2005] which emphasised the utility of the approach for multi-agent system construction and calibration. Here we focus on spatio-temporal data analysis and a real world application.

The framework, Forced Dynamical Pattern Discovery (FDPD), is introduced in Section 2. We then describe the application in Section 3. Results are presented and discussed in Section 3.4. We conclude and discuss future work in Section 4.

2 FORCED DYNAMICAL PATTERN DISCOVERY

2.1 Space, Time and Data

Data mining is the automated search for knowledge that is hidden in large collections of data. The primitives of such data collections are known as 'attributes'. In environmental science and other areas where space-time behaviour is a major focus of investigation, it is common to have many attributes in the data collection whose values change with space and time. Thus it is desirable to have a 'space-time-stamp' for these. One way to approach the mining of such data collections is to extend out from traditional data mining techniques, considering space and time simply as three (or four) additional dimensions to the n-dimensional data mining operations. Because attribute values tend to be highly correlated in space-time, the computational impact of these additional dimensions can be mitigated by clever choice of when to treat the attributes as dependent (with respect to their space-time separation).

Spatio-temporal problems that involve a relatively small number of variables (attributes) have been the focus of physical and mathematical modelling for many hundreds of years under the rubric of (what has become) Dynamical Systems Theory (DST). Therefore a second approach to spatio-temporal data mining is to consider the spatio-temporal behaviour as primary, and look at ways of incorporating many variables (attributes) in DST type models.

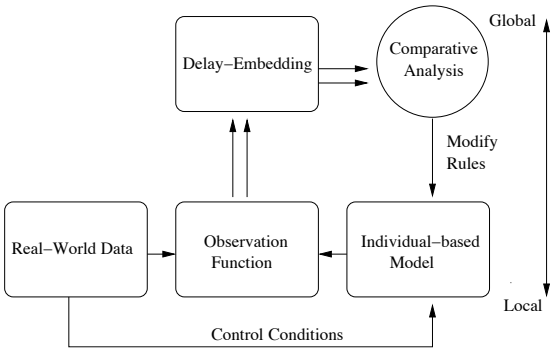


Figure 1: Framework overview.

DST provides a powerful tool, known as *delay-embedding*, for reconstructing the dynamical behaviour of a system from data as long as the system is deterministic and the complexity not too great. Recently, advances in this area have provided a rigorous foundation for reconstructing systems which are non-autonomous, that is ones that are forced (or driven) by stochastic or deterministic signals (e.g. noise). This is significant because it provides a way to treat higher dimensional systems in a modular fashion: as consisting of mutually forced subsystems. The framework we present here is based on this theoretical advance, which we explain in detail in the next section.

2.2 The Framework

Having motivated the style of approach at least partially, we present the essential components and architecture of the framework in Figure 1.[2] The *Individual-Based Model* (IBM) denotes some kind of complex system model; it will usually comprise a large number of interacting parts which update over time according to domain specific rules. The *observation function* approximates one or more spatially continuous variables from the output of the IBM and the real world data. Although complex systems are by definition high-dimensional, it is very often the case that their dynamical behaviour can be captured remarkably well in a much smaller number of new

[2]The complete framework includes a mechanism not shown in Figure 1 which provides a second level of feedback that is local rather than global. For details see Campbell et al. [2005].

dimensions, as long as those new dimensions are well chosen[3]. This principle is exploited to great effect by the theory of *delay-embedding*, and this underpins the global analysis end of the framework. Each of these components will be described in more detail below, but first we describe the flow around the diagram in Figure 1.

The real-world data provides a set of initial conditions to the IBM - one for every temporal 'snapshot' - which together we term *control conditions*. The IBM uses these control conditions as reference points, and to guide preliminary calibration of rules. The values of the components of the IBM over a 'run' of the model are a spatio-temporal data series themselves, ones which are usually of higher temporal (and possibly spatial) resolution than the empirical data. These two spatio-temporal data series, one real, one synthetic, are then both 'observed'. In data mining terms this can be thought of as playing an attribute-selection role where it may be necessary to construct a new attribute from the existing ones. Once a particular continuous variable has been obtained from the IBM and another for the real-world data according to the same observation function, they are both delay-embedded. The way in which this transforms an observation is subtle, but the upshot is it allows the comparison of the global behaviour of the real world data and the IBM in a space optimised for its ability to present the relevant dynamics in a smaller number of dimensions[4].

Because the IBMs' data series is of higher temporal resolution, the many space-time locations for which there is no empirical equivalent act as unknowns. Therefore there are potentially many ways (hypothetical scenarios) for the IBM data series to be different (due to different rules or a different model completely) and yet in 'agreement' with the empirical data. In embedding space these unknowns describe volumes - rather than the trajectories of observed data - and these may be bounded by (space-time) proximity to known observations and/or expert knowledge. The volumes represent 'potential trajectories': potential behaviours of the system. 'Agreement' requires that all synthetic data trajectories are coincident with the empirical ones, and within any volumes. The differences between the IBM and the empirical data in embedding space are used firstly to align (calibrate) the IBM with the empirical data, and then to explore hypotheses consistent with it. This is an iterative process.

[3]That is, they describe a good *basis*.

[4]There is a price to pay for this optimality: the delay-embedding space is without physical meaning. However, as we are using it to relate one embedded system to another, rather than in an absolute sense, this is not an issue.

Delay-Embedding. Also known as *geometry from a time series*, this mathematical fact provides a deep theoretical foundation for the analysis of time series generated by nonlinear deterministic dynamical systems. The profound insight of embedding is that an accessible variable can explicitly retrieve unseen internal degrees of freedom (Gershenfeld [1999]). In fact, under reasonable technical conditions it allows reconstruction of the *entire phase space behaviour* from a single scalar observable. Phase space is the collection of possible states (e.g. attributes) of the system that specify the system completely - they all we need to know to have complete knowledge of the immediate future.

As well as the phase space, which we denote M, a deterministic dynamical system is characterised by an update rule, f, which defines how the system changes over time. The rule dictates the new point in phase space, $\mathbf{x} \in M$, of the system, given the current point (or state). The next state defined as a function of the current one is known as an *image*. A rule must have a single value for each point (state) in phase space, but there could be several different states that give rise to the same image. In discrete time, we have

$$\mathbf{x}_{i+1} = f(\mathbf{x}_i). \tag{1}$$

which traces out a trajectory in phase space over time. Thus the temporal behaviour is *geometrised* - it becomes a static, but potentially complicated, geometrical object. Crucially, this geometry is unique to the dynamical behaviour, it is invariant to changes in the way the system is analysed. The general idea is that it is possible to reconstruct the phase-space and dynamical behaviour (trajectories in phase space), by considering delayed versions of a single scalar observable. The delays form a new space known as an *embedding* space, defined by the embedding map Φ. See Figure 6 for an example.

Takens' theorem (Takens [1981]) states that almost every smooth map $\Phi : M \to \mathbb{R}^d, d \geq 2m + 1$, is such an embedding, where m is the dimension of the phase space. Because $d \geq 2m + 1$ is a sufficient condition, an embedding may be possible for $m < d \leq 2m$. This is a remarkable result, however the assumptions that the system is deterministic, i.e. it has no random aspect, and autonomous, i.e. it is not 'driven' or 'forced' by any external events, are not justified for many real world applications. This is particularly true for the biological and environmental sciences, where a purely deterministic approach could justifiably be considered foolhardy. Therefore it is significant that there has been a number of generalisations of the the-

ory along these lines. A method for reconstructing input-output systems was conjectured in Casdagli [1992] and subsequently proved by Stark and colleagues (Stark [1999]). They present a number of forced embedding theorems, including the case of stochastic forcing (Stark et al. [2003]). This leads to models of the form

$$\mathbf{x}_{i+1} = f(\mathbf{x}_i, \omega_i) \tag{2}$$

where $\mathbf{x} \in M$ is a point in the state space and $\omega \in \Sigma$ is a stochastic forcing term. This dramatically increases the degree to which real world data can be considered from a formal basis.

In particular, it has led to (and in fact was partially motivated by) the case of spatially extensive systems. The basic idea, presented in Orstavik and Stark [1998], is that a spatially extended system can be modelled as a number of local spatial subsystems weakly coupled to, and driven by, the 'noise' at their boundaries. In other words M is a local spatial region, and it forms the part of the whole system from which we can take observables. Therefore our delays may also have a spatial component - this corresponds to recording values of neighbours. The forcing by the rest of the lattice is modelled similarly to the standard forced system above, except that the update of the forcing is now itself dependent on $\mathbf{x}$. This means that it is no longer possible to have an *embedding* in the rigorous, technical sense of the word. However, because the effect M has on the forcing dynamics is generally small, Orstavik and Stark felt that the delay embedding technique still provides the correct intuition. This is supported by experimental results in their paper, and in a number of other studies before and since; in fact it goes a long way to explaining the otherwise 'surprising' success of reconstruction techniques in such situations. The consequence of all this is a unified framework for dealing with high-dimensional dynamical systems, that has a holistic treatment of space-time. We have omitted many details, and direct the interested reader to the seminal papers mentioned.

Individual-based Modelling. This term is used here to encompass agent-based models, as well as cellular automata and other connectionist models. They all model spatially distributed individuals (or objects), and in the case of agent based systems, model them as decision makers with traits that interact and evolve in time. They are representative of the bottom-up approach to complex systems modelling - program the details and the macroscopic patterns will emerge. Thus they are ideal for for inputting domain knowledge and exploring hypothetical scenarios. A major goal of this work is to provide a global spatio-temporal representation of these knowledge rich models which captures the essentials of their dynamical behaviour. As the delay-embedding space is not scale dependent, this will facilitate transfer of knowledge.

The Observation Function. Because IBMs are typically discrete state models, and delay-embedding is only theoretically well defined for continuous state systems, we need to obtain this continuous variable somehow. This will be problem dependent, and because of universality of the delay-embedding approach (any variable can be used) there is a certain amount of freedom in the way it is done. However, this generally consists of the following steps: identifying the target physical variable; considering the most appropriate continuous version of it; and constructing an approximation to this (possibly hypothetical) continuous variable. A rigorous formulation of this device may be possible through measure theory, however at this stage we will rely on the approximations above, and leave examination of this issue for future work.

3 APPLICATION: TRANSIT-ORIENTED DEVELOPMENTS

Transit Oriented Developments (TODs) (Gilbert and Ginn [2001]) are one proposed solution to the problems of urban sprawl, air-pollution, traffic congestion and many others that characterise most large cities. The basic principle of a TOD is a mixed-use community within an average walking distance of a transit stop core commercial area. Given relevant data, and some domain knowledge, a mathematical model that explores this concept can be extremely useful. This research seeks to build and train a model with a transparent architecture that can be used to better understand existing transit-centric urban patterns in space and time, and simultaneously to explore TOD scenarios in a virtual sense using this knowledge. The simultaneity will be achieved using FDPD to provide a feedback loop between the data-driven modelling and the hypothesis exploration. In this paper however, we simply concentrate on a phase-space reconstruction of a simple variation of the empirical data. The goal is to look for evidence of relatively low-dimensional behaviour; this is discussed more in Section 3.4.

3.1 Data

The data consists of the number of dwellings in geospatial regions known as Census Districts (CDs)

Figure 2: CD Boundaries (P.I.F.U. [2005]).

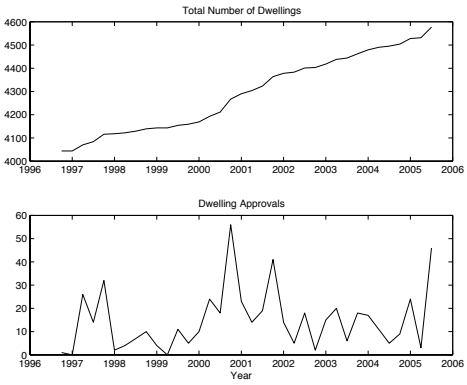


Figure 3: Aggregate statistics for Albion.

within a one-kilometre radius of train stations in the city of Brisbane, Australia. As this is a feasibility study of the method of analysis, rather than a study of the phenomena, this can be accomplished using data pertaining to one train station only. Figure 2 shows the one-kilometre radius around Albion train station. The number of dwelling approvals is available for each CD per quarter from September 1996 to June 2005, giving 36 temporal data points. Spatially there are 22 distinct CDs. Aggregate statistics (across all CDs) are in Figure 3.

3.2 Individual Based Model

The model we have created here is simply a higher spatial resolution version of the empirical data. The region surrounding the train stations was tesselated into a grid of $256 * 256$ cells, such that the number of cells in the most densely populated CD was just greater than the number of dwellings. Cells were then initialised as being a dwelling or not probabilistically, such that in the limit of infinite sized regions, the absolute number of dwellings in that CD is equal to the empirical data. Rather than

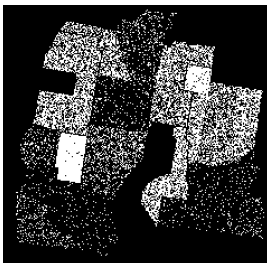


Figure 4: Starting configuration for the IBM.

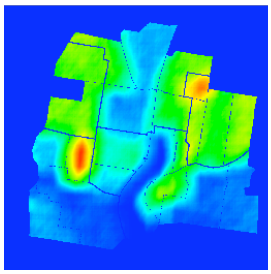


Figure 5: Spatially continuous dwelling density.

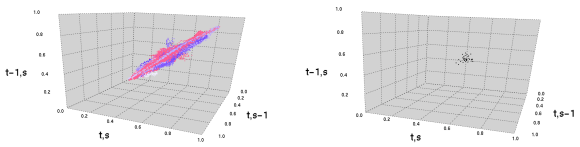
knowledge-based transition rules, this initialisation is then carried out again for each successive empirical time-series record. Figure 4 shows a spatial snapshot for the first time-series record.

3.3 Observation Function

The first thing we have to do is define our target variable. Here we will use dwelling spatial density, as this can be defined everywhere (it is continuous in space and time) and we have strongly relevant empirical support for this from which to construct the observable. An approximation to the dwelling density is found by counting the number of dwellings within a region of a certain number of cells, surrounding the centroid cell at which the observable is being estimated, and dividing by the total number of cells in that region. This acts like a smoothing operation. In the results below, a $27*27 = 729$ region of cells was used. The size of the region roughly corresponds to the degree of spatial smoothing. Close to the boundaries, densities were calculated by reducing the size of the region. The spatial delay, referred to in the 'delay-embedding' part of section 2.2, was taken only along one spatial dimension as the system is rotationally symmetric with respect to distance from a train station. These distances were calculated all over the grid (including from other train stations than Albion) and spatial delays were always taken in the direction of greatest distance. The spatial delay in the results below was 5 cells.

3.4 Results and Discussion

Figure 6a displays an embedding space for measurements taken at every spatial location, Figure 6b for only a single site. The vertical axis corresponds to measurements with one temporal lag (three months), the two horizontal axes correspond to measurements 'now' and 'now' with one spatial lag (5 pixels) respectively from left to right. According to Taken's theorem, if a sufficient number



(a) Whole system (b) Single site

Figure 6: Embedding spaces

of delays are used, the points should lie on a surface in embedding space, otherwise they will fall in a volume. High dimensional and/or noisy systems will have points evenly distributed in that volume; systems for which the number of delays is insufficient to fully 'unfold' the attractor, but which still have relatively low-dimensional signals will have a large amount of structure present in the volume. In Figure 6a, different colours represent different 'blocks' of measurements; a similar colour indicates close proximity in space and/or time and highlights the amount of structure present in the embedding. Obviously 3 delays are insufficient to embed the system, but the shape of the cloud of points is strongly suggestive of structure in low-dimensional deterministic signals. The only reason for the 3 delay limitation here is visualisation, the quantitative components of the framework are not so limited.

A second general feature of the cloud of points is it hugs the $x = y = z$ axis. This is usually a sign that the size of the delay should be increased, however the inadequacy of 3 dimensions for the embedding is more fundamental. Currently, work on the CWM component of the framework is determining the number of dimensions required to fully unfold most of the dynamics of the system. In terms of domain-relevant features, there is an interesting asymmetry between spatio-temporal behaviour below about 0.5 dwellings per cell and above it. There is also quite anomalous spatio-temporal behaviour for two small ranges of density represented by the two lobes protruding from above and below the main cloud. Identifying such characteristics can then guide creation of IBMs with higher spatio-temporal resolution, which is work in progress.

4 Conclusion and Future Work

We have presented Forced Dynamical Pattern Discovery as a framework for spatial data analysis and hypothesis exploration. We have shown how the framework can be applied to a real world problem and given some preliminary results for the phase-space reconstruction component. The strong signal of low-dimensional structure is promising. Future work will focus mainly on two aspects: exploring a measure theoretic version of the measuring device, and more sophisticated individual-based modelling. Systematising interpretation of shapes and features in embedding space will also be investigated. Implementation-wise, there is the issue of computational complexity, which is currently being addressed by use of a graphics processing unit (GPU) for the simulations. We will move more computation to the GPU, and further exploit its visualisation capabilities.

References

Campbell, A., B. Pham, and Y.-C. Tian. Delay-embedding approach to multi-agent system construction and calibration. In Zerger., A. and Argent, R., editors, *MODSIM 2005 International Congress on Modelling and Simulation*, Modelling and Simulation Society of Australia and New Zealand, pages 113–119, 2005.

Casdagli, M. A dynamical systems approach to modeling input-output systems. In Casdagli, M. and Eubank, S., editors, *Nonlinear Modeling and Forecasting*. Addison-Wesley, Santa-Fe Institute, 1992.

Chawla, S. An invitation to spatio-temporal data mining: Definition and applications. 2003.

Gershenfeld, N. *The nature of mathematical modeling*. Cambridge University Press, Cambridge, United Kingdom, 1999.

Gilbert, D. and S. Ginn. Transit oriented sustainable developments. Technical report, Built Environment Research Unit, Building Division, Queensland Department of Public Works, 2001.

Huang, H.-C. and N. Cressie. Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics and Data Analysis*, 22:159–175, 1996.

Kantz, H. and T. Schrieber. *Nonlinear Time Series Analysis*. Cambridge University Press, 1997.

Orstavik, S. and J. Stark. Reconstruction and cross-prediction in coupled map lattices using spatio-temporal embedding techniques. *Physics Letters A*, 247(1-2):145 – 160, 1998.

P.I.F.U. Census district maps, 2005. Planning and Information Forecasting Unit, Department of Local Government and Planning, QLD.

Stark, J. Delay embeddings for forced systems. i. deterministic forcing. *Journal of Nonlinear Science*, 9(3): 255 – 332, 1999.

Stark, J., D. Broomhead, M. Davies, and J. Huke. Delay embeddings for forced systems. ii. stochastic forcing. *Journal of Nonlinear Science*, 13(6):519 – 577, 2003.

Takens, F. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics*, 898, 1981.