3.2   Sampling

Chapter 2, on P. 130, we discussed the problem of making a uniformly gridded map from irregularly spaced observations. But not any set of observations proves adequate to the purpose. The most fundamental problem generally arises under the topic of "sampling" and "sampling error." This subject is a large and interesting one in its own right,[60] and we can only outline the basic ideas.

The simplest and most fundamental idea derives from consideration of a one-dimensional continuous function, $f(q)$, where q is an arbitrary independent variable, usually either time or space, and $f(q)$ is supposed to be sampled uniformly at intervals, Δq, an infinite number of times to produce the infinite set of sample values $\{f(n\Delta q)\}$, $-\infty \leq n \leq \infty$. The sampling theorem, or sometimes the "Shannon-Whittaker Sampling Theorem"[61] is a statement of the conditions under which $f(q)$ should be reconstructable from the sample values. Let the Fourier transform of $f(q)$ be defined as

{36023}
$$\hat{f}(r) = \int_{-\infty}^{\infty} f(q)e^{2i\pi rq}dq\,, \tag{3.14}$$

and assumed to exist. The sampling theorem asserts that a necessary and sufficient and assumed to exist condition to perfectly reconstruct $f(q)$ from its samples is that,

{36024}
$$|\hat{f}(r)| = 0\,, \quad |r| \geq 1/(2\Delta q)\,. \tag{3.15}$$

From the theorem emerges the Shannon-Whittaker formula for the reconstruction,

{36025}
$$f(q) = \sum_{n=-\infty}^{\infty} f(n\Delta q)\frac{\sin[(2\pi/2\Delta q)(q - n\Delta q)]}{(2\pi/2\Delta q)(q - n\Delta q)}\,. \tag{3.16}$$

Mathematically, the Shannon-Whittaker theorem is surprising—because it provides a condition under which a function at an uncountable infinity of points—the continuous line—can be perfectly reconstructed from information known only at a countable infinity, $n\Delta q$, of them. For present purposes, an intuitive interpretation is all we seek and this is perhaps best done by considering a special case in which the conditions of the theorem are violated.

Figure 3.2 displays an ordinary sinusoid whose Fourier transform can be represented as

{36026}
$$\hat{f}(r) = \tfrac{1}{2}(\delta(r - r_0) - \delta(r + r_0))\,, \tag{3.17}$$

which is sampled as depicted, and in violation of the sampling theorem. It is quite clear that there is at least one more perfect sinusoid, the one depicted with the dashed line, which is completely consistent with all the sample points and which cannot be distinguished from it using the measurements alone. A little thought shows that the apparent frequency of this new sinusoid is,

{36027}
$$r_a = r_0 \pm \frac{n}{\Delta q} \tag{3.18}$$

such that

{36028}
$$|r_a| \leq \frac{1}{2\Delta q}\,. \tag{3.19}$$

The samples cannot distinguish the true high frequency sinusoid from this low frequency one, and the high frequency can be said to masquerade or "alias" as the lower frequency one.[62] The
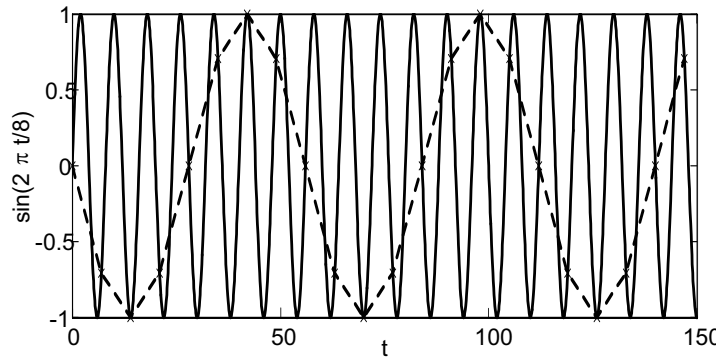
Figure 3.2: Effects of undersampling a periodic function: solid curve is $y(t) = \sin(2\pi t/8)$ sampled at time intervals of $\Delta t = 0.1$. The dashed curve is the same function, but sampled at intervals $\Delta t = 7$. With this undersampling, the curve of frequency $s = 1/8$ time units , is aliased into one that appears to have a periodicity of period $s_a = 1/8 - 1/7 = 1/56 < 1/14$. That is, the aliased curve appears to have a period of 56 time units.

Fourier transform of a sampled function is easily seen to be periodic with period $1/\Delta q$ in the transform domain, that is, in the r space.[63] Because of this periodicity, there is no point in computing its values for frequencies outside $|r| \leq 1/2\Delta q$ (we make the convention that this "baseband," i.e., the fundamental interval for computation, is symmetric about $r = 0$, over a distance $1/2\Delta q$; see fig. 3.3). Frequencies of absolute value larger than $1/2\Delta q$, the so-called Nyquist frequency, cannot be distinguished from those in the baseband, and alias into it. Fig. 3.3 shows a densely sampled, non-periodic function and its Fourier transform compared to that obtained from the undersampled version overlain. Undersampling is a very unforgiving practice.

The consequences of aliasing range from the negligible to the disastrous. A simple example is that of the principal lunar tide, usually labelled M_2, with a period of 12.42 hours, $r = 1.932$ cycles/day. An observer measures the height of sea level at a fixed time, say 10 AM. each day so that $\Delta q = 1$ day. Applying the formula (3.18), the apparent frequency of the tide will be .0676 cycles/day for a period of about 14.8 days ($n = 2$). To the extent that the observer understands what is going on, she will not conclude that the principal lunar tide has a period of 14.8 days, but will realize that the true period can be computed through (3.18) from the apparent one. But without that understanding, some bizarre theory might be produced.[64]

The reader should object that the Shannon-Whittaker theorem applies only to an infinite number of perfect samples and that one never has either perfect samples or an infinite number
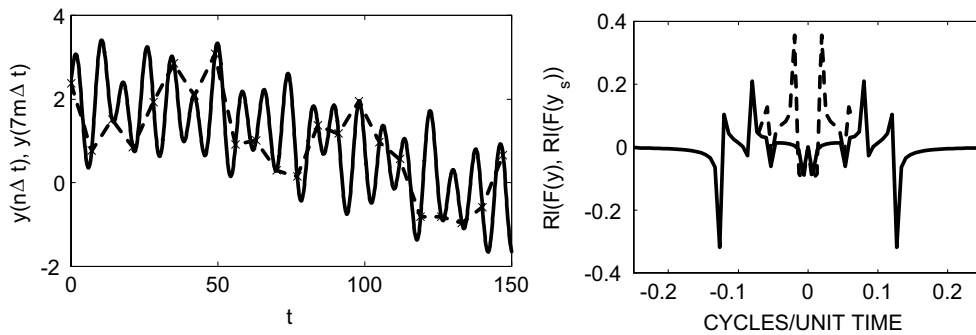
Figure 3.3: (Left panel), a non-periodic function sampled at intervals $\Delta t = 0.1$, and the same function sampled at intervals $\Delta t = 7$ time units. (Right panel). Real part of the Fourier components of the two functions shown in the left panel. The subsampled function has a Fourier transform confined to $|s| \leq 1/(2 \cdot 7)$ while that of the original, more densely sampled function extends to $|s| \leq 1/0.1 = 10$, most of which is not displayed. Note that the subsampled function has a very different Fourier transform from that of the original densely sampled one. Both transforms are periodic in frequency s, with period equal to the width of their corresponding basebands. (This periodicity is suppressed in the plot.) Note in particular how erroneous an estimate of the temporal derivative of the undersampled function would be in comparison to the highly sampled one.

ampgenl.eps}

of them. In particular, it is true that if the duration of the data in the q domain is finite, then it is impossible for the Fourier transform to vanish over any finite interval, much less the infinite interval above the Nyquist frequency.[65] Nonetheless, the rule of thumb that results from (3.16) has been found to be quite a good one. The deviations from the assumptions of the theorem are usually dealt with by asserting that sampling should be done so that,

$$\Delta q \ll 1/2r_0 \,. \tag{3.20}$$

{36029}

(The example in Fig. 3.3 demonstrates that Fourier transforms of finite duration sampled signals are nonetheless useful.) Many extensions and variations of the sampling theorem exist—taking account of the finite time duration , the use of "burst-sampling" and known function derivatives, etc.[66] Most of these variations are sensitive to noise. There are also extensions to multiple dimensions,[67] which are required for mapmaking purposes. Because failure to acknowledge the possibility that a signal is undersampled is so dire, one concludes that consideration of sampling is critical to any discussion of field data.

3.2.1 One-Dimensional Interpolation

Let there be two observations $[y_1, y_2]^T = [x_1+n_1, x_2+n_2]^T$ located at positions $[r_1, r_2]^T$ where n_i are the observation noise. We require an estimate of $x(\tilde{r})$, where $r_1 < \tilde{r} < r_2$. The formula (3.16) is unusable—there are only two noisy observations, not an infinite number of perfect ones. We could try using linear interpolation:

$$\tilde{x}(\tilde{r}) = \frac{|r_2 - \tilde{r}|}{|r_2 - r_1|}\, y(r_1) + \frac{|r_1 - \tilde{r}|}{|r_2 - r_1|}\, y(r_2) \,. \tag{3.21}$$

{36030}

If there are N data points, r_1, $1 \leq i \leq N$, then another possibility is Aitken-Lagrange interpolation:[68]

$$\tilde{x}(\tilde{r}) = \sum_{j=1}^{M} l_j(\tilde{r}) y_j \,, \tag{3.22}$$

{36031a}

$$l_j(\tilde{r}) = \frac{(\tilde{r} - r_1)\cdots(\tilde{r} - r_M)}{(r_j - r_1)\cdots(r_j - r_M)} \,. \tag{3.23}$$

{36031b}

Eqs. (3.21)-(3.23) are only two of many possible interpolation formulas. When would one be better than the other? How good are the estimates? To answer these questions, let us take a different tack, and employ the Gauss-Markov theorem, assuming we know something about the necessary covariances.

Suppose either $\langle x \rangle = \langle n \rangle = 0$ or that a known value has been removed from both (this just keeps our notation a bit simpler). Then,
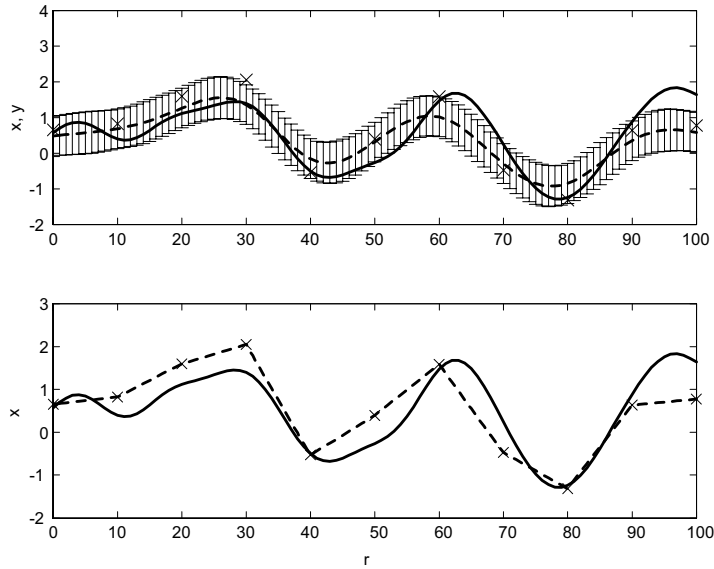
Figure 3.4: (Both panels) Solid curve is the "true" curve, $x(r)$, from which noisy samples (denoted 'x') have been obtained. $x(r)$ was generated to have a true covariance $S = \exp\left(-r^2/100\right)$ and the "data", $y(r_i) = x(r_i) + n_i$ where $\langle n_i \rangle = 0$, $\langle n_i n_j \rangle = (1/4)\,\delta_{ij}$, the values generated from a Gaussian probability density. In the lower panel, linear interpolation is used to generate the estimated values of $x(r)$ (dashed line). The estimates are identical to the observations at $r = r_i$. In the upper panel, objective mapping was used to make the estimates (dashed line). Note that $\tilde{x}(r_i) \neq y(r_i)$, and that an error bar is available—as plotted. The true values are generally within one standard deviation of the estimated value (but about 35% of the estimated values would be expected to lie outside the error bars), and the estimated value is within two standard deviations of the correct one everywhere. The errors in the estimates, $\tilde{x}(r_i) - x(r_i)$ are clearly spatially correlated, and can be inferred from Eq. 3.28 (not shown). The values of $x(r)$ were generated to have the inferred covariance S, by forming the matrix, $\mathbf{S} = \text{toeplitz}\left(S(r_i, r_j)\right)$, and obtaining its symmetric factorization , $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$. $\mathbf{x}(r) = \mathbf{U}\boldsymbol{\Lambda}\boldsymbol{\alpha}$, where the elements of $\boldsymbol{\alpha}$ are pseudo-random numbers.

{interp1.eps}

$$\mathbf{R}_{xy}(\tilde{r}, r_j) \equiv \langle x(\tilde{r})y(r_j)\rangle = \langle x(\tilde{r})(x(r_j) + n(r_j))\rangle = \mathbf{R}_{xx}(\tilde{r}, r_j), \tag{3.24}$$

$$\mathbf{R}_{yy}(r_i, r_j) \equiv \langle (x(r_i) + n(r_i))(x(r_j) + n(r_j))\rangle \tag{3.25}$$

$$= \mathbf{R}_{xx}(r_i, r_j) + \mathbf{R}_{nn}(r_i, r_j), \tag{3.26}$$

where it has been assumed that $\langle x(r)n(q)\rangle = 0$.

From (2.394), the best linear interpolator is

$$\tilde{\mathbf{x}} = \mathbf{B}\mathbf{y}, \qquad \mathbf{B}(\tilde{r}, \mathbf{r}) = \sum_{j=1}^{M} \mathbf{R}_{xx}(\tilde{r}, r_j) \left\{ \mathbf{R}_{xx} + \mathbf{R}_{nn} \right\}_{ji}^{-1} \tag{3.27} \quad \text{\{36033a\}}$$

($\left\{ \mathbf{R}_{xx} + \mathbf{R}_{nn} \right\}_{ji}^{-1}$ means the ji element of the inverse matrix) and the minimum possible error which results is

$$\mathbf{P}(\tilde{r}_\alpha, \tilde{r}_\beta) = \mathbf{R}_{xx}(\tilde{r}_\alpha, \tilde{r}_\beta) - \sum_{j}^{M}\sum_{i}^{M} \mathbf{R}_{xx}(\tilde{r}_\alpha, r_j)\{\mathbf{R}_{xx} + \mathbf{R}_{nn}\}_{ji}^{-1}\mathbf{R}_{xx}(r_i, \tilde{r}_\beta) \tag{3.28} \quad \text{\{36033b\}}$$

(here $\mathbf{R}_{xx}$, $\mathbf{P}(\tilde{r}, \tilde{r})$ are both scalars), and $\tilde{\mathbf{n}} = \mathbf{y} - \tilde{\mathbf{x}}$.

Results for both linear interpolation and objective mapping are shown in Fig. 3.4. Notice that like other interpolation methods, the optimal one is simply a linear combination of the data. If any other set of weights $\mathbf{B}$ is chosen, then the interpolation is not as good as it could be in the mean-square error sense; the error of any such scheme can be obtained by substituting it into (2.391) and evaluating the result (the true covariances still need to be known).

Looking back now at the two familiar formulas (3.21, 3.22), it is clear what is happening: they represent a choice of $\mathbf{B}$. Unless the covariance is such as to produce one of the two sets of weights as the optimum choice, neither Aitken-Lagrange nor linear (nor any other common choice, like a spline) is the best one could do. Alternatively, if either of (3.21), (3.22)–(3.23) was thought to be the best one, they are equivalent to specifying the solution and noise covariances.

If interpolation is done for two points, $\tilde{r}_\alpha$, $\tilde{r}_\beta$, the error of the two estimates will usually be correlated, and represented by $\mathbf{P}(\tilde{r}_\alpha, \tilde{r}_\beta)$. Knowledge of the correlations between the errors in different interpolated points is often essential—for example, if one wishes to interpolate to uniformly spaced grid points so as to make estimates of derivatives of x. Such derivatives might be numerically meaningless if the mapping errors are small scale (relative to the grid spacing) and of large amplitude. But if the mapping errors are large scale compared to the grid, the derivatives may tend to remove the error and produce better estimates than for x itself.

Both linear and Lagrangian weights will produce estimates which are exactly equal to the observed values if $\tilde{r}_\alpha = r_p$, that is, on the data points themselves. Such a result is characteristic

of "true interpolation." If no noise is present, then the observed value is the correct one to use at a data point. In contrast, the Gauss-Markov estimate will differ from the data values at the data points, because the estimator attempts to reduce the noise in the data by averaging over all observations, not just the one at the local point. The Gauss-Markov estimate is thus not a true interpolator; it is instead a "smoother." One can recover true interpolation if $\|\mathbf{R}_{nn}\| \to 0$, although the matrix being inverted in (3.27), (3.28) can become singular. The weights $\mathbf{B}$ can be fairly complicated if there is any structure at all in either of $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$. The estimator takes explicit account of the expected spatial structure of both $\mathbf{x}$, $\mathbf{n}$ to weight the data in such a way as to most effectively "kill" the noise relative to the signal. One is guaranteed that no other linear filter can do better.

If $\|\mathbf{R}_{nn}\| \gg \|\mathbf{R}_{xx}\|$, $\tilde{\mathbf{x}} \to \mathbf{0}$, manifesting the bias in the estimator; this bias was deliberately introduced so as to minimize the uncertainty (minimum variance about the true value). Thus, interpolated values tend toward zero, particularly far from the data points. For this reason, it is common to use expressions such as (2.411) to first remove the mean, prior to mapping the residual, adding the estimated mean back in afterward. The interpolated values of the residuals are nearly unbiased, because their true mean is nearly zero. Rigorous estimates of $\mathbf{P}$ for this approach require some care, as the mapped residuals contain variances owing to the uncertainty of the estimated mean,[69] but the corrections are commonly ignored.

As we have seen, the addition of small positive numbers to the diagonal of a matrix usually renders it non-singular. In the formally noise-free case, $\mathbf{R}_{nn} \to \mathbf{0}$, and one has the prospect that $\mathbf{R}_{xx}$ by itself may be singular. To understand the meaning of this situation, consider the general case, involving both matrices. Then the symmetric form of the SVD of the sum of the two matrices is,

$$\mathbf{R}_{xx} + \mathbf{R}_{nn} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T. \tag{3.29}$$

If the sum covariance is positive definite, $\mathbf{\Lambda}$ will be square with $K = M$ and the inverse will exist. If the sum is not positive definite, but is only semi-definite, one or more of the singular values will vanish. The meaning is that there are *possible* structures in the data which have been assigned to neither the noise field nor the solution field. This situation is realistic only if one is truly confident that $\mathbf{y}$ does not contain such structures. In that case, the solution,

$$\tilde{\mathbf{x}} = \mathbf{R}_{xx}(\mathbf{R}_{xx} + \mathbf{R}_{nn})^{-1}\mathbf{y} = \mathbf{R}_{xx}(\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T)\mathbf{y}, \tag{3.30}$$

will have components of the form $0/0$, the denominator corresponding to the zero singular values and the numerator to the absent, impossible, structures of $\mathbf{y}$. One can arrange that the ratio of these terms should be set to zero (e.g., by using the SVD). But such a delicate balance is not necessary. If one simply adds a small white noise covariance to $\mathbf{R}_{xx} + \mathbf{R}_{nn} \to \mathbf{R}_{xx} + \mathbf{R}_{nn} + \epsilon^2\mathbf{I}$,

or $\mathbf{R}_{xx} \to \mathbf{R}_{xx} + \epsilon^2 \mathbf{I}$, one is assured, by the discussion of tapering, that the result is no longer singular—all structures in the field are being assigned either to the noise or the solution (or in part to both).

Anyone using a Gauss-Markov estimator to make maps must do checks that the result is consistent with the prior estimates of $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$. Such checks include determining whether the difference between the mapped values at the data points and the observed values have numerical values consistent with the assumed noise variance; a further check involves the sample autocovariance, of $\tilde{\mathbf{n}}$ and its test against $\mathbf{R}_{nn}$ (see books on regression for such tests). The mapped field should also have a variance and covariance consistent with the prior estimate $\mathbf{R}_{xx}$. If these tests are not passed, the entire result should be rejected.

3.2.2 Higher Dimensional Mapping

We can now immediately write down the optimal interpolation formulas for an arbitrary distribution of data in two or more dimensions. Let the positions where data are measured be the set $\mathbf{r}_j$ with measured value $\mathbf{y}(\mathbf{r}_j)$, containing noise $\mathbf{n}$. It is assumed that aliasing errors are unimportant. The mean value of the field is first estimated and subtracted from the measurements and we proceed as though the true mean were zero.[70] Fundamentally, it is nothing more than an application of the Gauss-Markov theorem in two (most commonly) dimensions.[71]

One proceeds exactly as in the case where the positions are scalars, minimizing the expected mean-square difference between the estimated and the true field $\mathbf{x}(\tilde{\mathbf{r}}_\alpha)$. The result is (3.27), (3.28), except that now everything is a function of the vector positions. If the field being mapped is also a vector (e.g., two components of velocity) with known covariances between the two components, then the elements of $\mathbf{B}$ become matrices. The observations could also be vectors at each point.

An example of a two-dimensional map is shown in figure 3.5:. The "data points", $y(\mathbf{r}_i)$, are the x-s, while estimates of $\tilde{x}(\mathbf{r}_i)$ on the uniform grid were wanted. The a priori noise was set to $\langle \mathbf{n} \rangle = 0$, $\mathbf{R}_{nn} = \langle n_i n_j \rangle = \sigma_n^2 \delta_{ij}$, $\sigma_n^2 = 1$, and the true field covariance was $\langle \mathbf{x} \rangle = 0$, $\mathbf{R}_{xx} = \langle \mathbf{x}(\mathbf{r}_i)\mathbf{x}(\mathbf{r}_j) \rangle = P_0 \exp -|\mathbf{r}_i - \mathbf{r}_j|^2/L_2$, $P_0 = 25$, $L_2 = 100$. Figure 3.5 also shows the estimated values and Figs. 3.5, 3.6 the error variance estimate of the mapped values. Notice that far from the data points, the estimated values are 0 : the mapped field goes asymptotically to the estimated true mean, with the error variance rising to the full value of 25, which cannot be exceeded. That is to say, when we are mapping far from any data point, the only real information available is provided by the prior statistics—that the mean is 0, and the variance about that mean is 25. So the expected uncertainty of the mapped field, in the absence of data, cannot exceed the prior estimate of how far from the mean the true value is likely to be. The
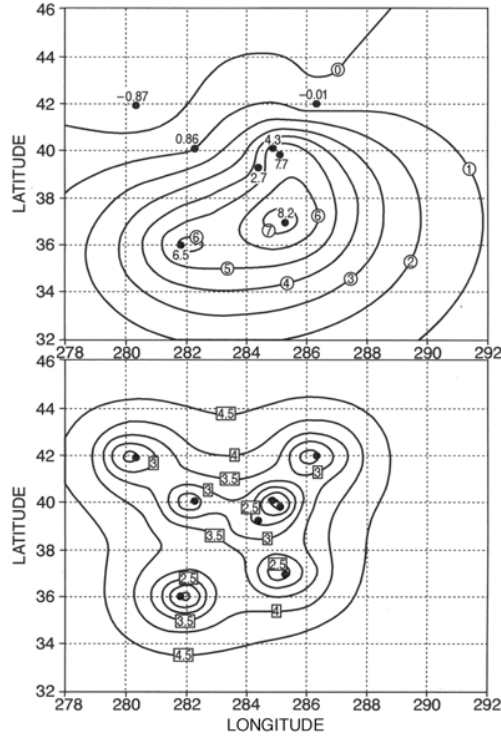
Figure 3.5: (Upper) Observations, shown as solid dots, from which a uniformly gridded map is desired. Contours were constructred using a fixed covariance and the Gauss-Markov estimate (3.27). Noise was assumed white with a variance of 1. (Lower) Expected standard error of the mapped field in the top panel. Values tend, far from the observations points, to a variance of 25, which was the specified field variance, and hence the largest expected error is $\sqrt{25}$. Note the minima centered on the data points.   {ocip3_15ab.ti

best estimate is then the mean itself.

A complex error structure of the mapped field exists—even in the vicinity of the data points. Should a model be "driven" by this mapped field, one would need to make some provision in the model accounting for the spatial change in the expected errors of this forcing.

In practice, most published objective mapping (often called "OI" for "objective interpolation," although we as we have seen, it is not true interpolation) has been based upon simple analytical statements of the covariances $\mathbf{R}_{xx}$, $\mathbf{R}_{nn}$ as used in the example: that is, they are commonly assumed to be spatially stationary and isotropic (depending upon $|\mathbf{r}_i - \mathbf{r}_j|$ and not upon the two positions separately nor upon their orientation). The use of analytic forms removes the necessity for finding, storing, and computing with the potentially very large $M \times M$ data
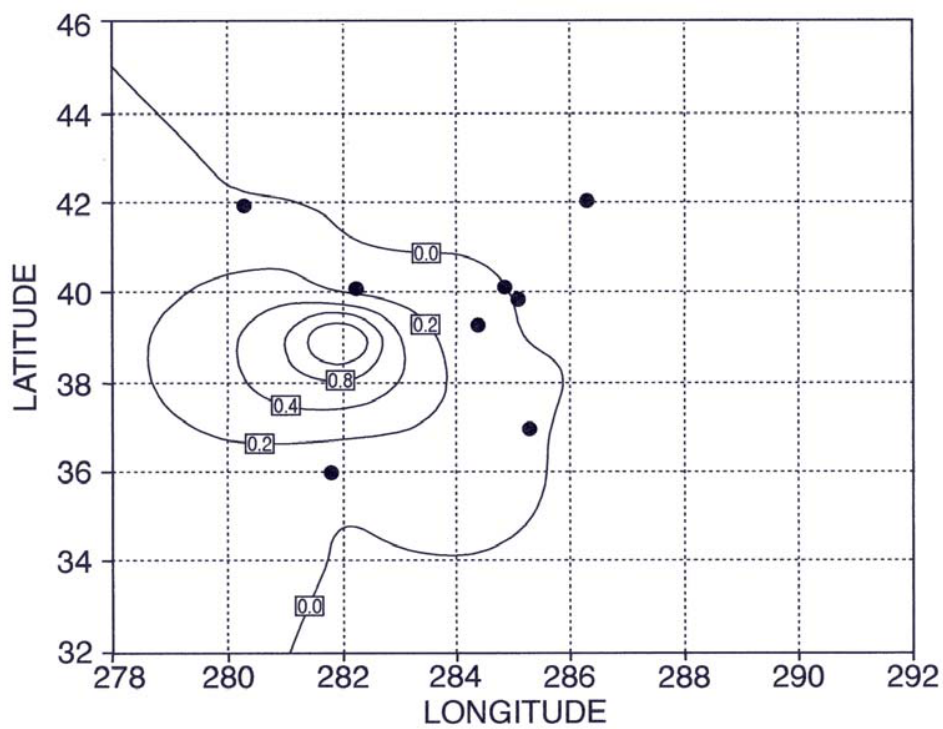
Figure 3.6: One of the rows of **P** corresponding to the grid point in Fig. 3.5 at 39°N, 282°E. Displays the expected correlations that occur in the errors of the mapped field. These errors would be important e.g., in any use that differentiated the mapped field. For plotting purposes, the variance was normalized to 1.

covariance matrices in which hypothetically every data or grid point has a different covariance with every other data or grid point. But the analytical convenience often distorts the solutions, as many fluid flows and other fields are neither spatially stationary nor isotropic.[72]

3.2.3   Mapping Derivatives

A common problem in setting up fluid models is the need to specify the fields of quantities such as temperature, density, etc., on a regular model grid. One also often must specify the derivatives of these fields for use in advection-diffusion equations,

{36036}
$$\frac{\partial C}{\partial t} + \mathbf{v} \cdot \nabla C = K \nabla^2 C, \tag{3.31}$$

where C is any scalar field of interest. Suppose one wished to estimate a spatial derivative as a one-sided difference,

{36037}
$$\frac{\partial C(\tilde{r}_1)}{\partial r} \sim \frac{C(\tilde{r}_1) - C(\tilde{r}_2)}{\tilde{r}_1 - \tilde{r}_2}. \tag{3.32}$$

Then one might think simply to subtract the two estimates made from eq. (3.27), producing

{36038}
$$\frac{\partial C(\tilde{r}_1)}{\partial r} \sim \frac{1}{\Delta r}(\mathbf{R}_{xx}(\tilde{r}_1, r_j) - \mathbf{R}_{xx}(\tilde{r}_2, r_j))(\mathbf{R}_{xx} + \mathbf{R}_{nn})^{-1}\mathbf{y}. \tag{3.33}$$

Alternatively, suppose we tried to estimate $\partial C/\partial r$ directly from (2.390), using $\mathbf{x} = C(r_1) - C(r_2)$. $\mathbf{R}_{yy} = \mathbf{R}_{xx} + \mathbf{R}_{nn}$, which describes the data, does not change. $\mathbf{R}_{xy}$ does change:

{36039}
$$\mathbf{R}_{xy} = \langle (C(\tilde{r}_1) - C(\tilde{r}_2))(C(r_j) + n(r_j)) \rangle = \mathbf{R}_{xx}(\tilde{r}_1, r_j) - \mathbf{R}_{xx}(\tilde{r}_2, r_j), \tag{3.34}$$

which when substituted into (2.394) produces (3.33). *Thus, the optimal map of the finite difference field is simply the difference of the mapped values.* More generally, the optimally mapped value of any linear combination of the values is that linear combination of the maps.[73]