# Chapter 4

# Field data analysis

The data analyzed in this work have all been recorded with the RAP (RAM-apparat) data loggers developed in Göttingen by *Steveling and Leven* [1992]. To measure the temporal variation of the electrical field, unpolarizable Ag-KCl electrodes (built at the GFZ Potsdam, based on a design from *Filloux* [1973]) were buried some decimeters below the surface, about 100 m pairwise apart from each other (E-W, N-S). Its membranes were taped with a handful of wet bentonite to ensure low and stable contact resistances. Fluxgate magnetometers from MAGSON company were deployed as three-component sensors for the magnetic variations. All instruments were oriented in a geomagnetic coordinate system, which is just slightly deviated from the geographic system in the northern study area (about one degree) what can be neglected. The declination in South Chile, however, amounts to roughly plus nine degrees and thus certainly has to be considered. The deviations were derived from the International Geomagnetic Reference Field (IGRF 2000, *Mandea and Macmillan* [2000]) and manually verified in the field.
Individual recordings with a sampling rate of two seconds usually covered data for about one week — a total runtime of four weeks was in most cases sufficient to get reasonable data quality at long periods. The target periods for the processing range from 10 s to $\sim 25 \cdot 10^3$ s, with eight frequencies per decade. Geomagnetic daily variations are not analyzed here (this was performed by *Friedel* [1997], with data from a subset of stations from the N Chile campaign in 1995). Data from the campaigns in 1993, '95 and '97 had to be completely rehandled for the multivariate analysis. At this opportunity, also local transfer functions were recalculated for the data sets from 1993 and '95.

Essential parts of this section concerning the multivariate analysis of data from the Central Andes are published in *Soyer and Brasse* [2001].

## 4.1 Processing techniques

In this thesis, both bivariate and multivariate data analysis methods are employed. Before the derivation of transfer functions with either technique, data were cleaned from isolated outliers in time domain using a running median filter. Then, within a 7-level cascaded decimation-scheme of decimation factor 4, data were pre-whitened by an adaptive, autoregressive filter and windowed to a length of 128 samples for the Fourier transformation (program package from *Egbert and Booker* [1986]).

### Bivariate processing

Within the bivariate analysis, which has only been used to calculate local magnetotelluric and geomagnetic transfer functions, a complex output quantity $Z$ (here: $E_x$, $E_y$ and $B_z$) is regarded as a linear combination of two error-free complex input quantities $X$, $Y$ (here: $B_x$, $B_y$) plus an inevitable misfit or error $e_i$, the difference between observation and prediction:

$$Z_i = aX_i + bY_i + e_i \tag{4.1}$$

Whereas the values of these quantities, i.e. the Fourier amplitudes of the respective field coordinates for the given time segment $i$, vary strongly in time, the complex coefficients $a$, $b$ – the transfer functions – are taken to be constants, which means in terms of the geomagnetic induction process, that the source field is supposed not to vary in geometry. Assume we have $N$ time segments, which are regarded as $N$ independent, identically distributed realizations of the experiment. Then the standard least squares approach to solve this problem leads to (e.g. *Sims et al.* [1971]; *Schmucker and Weidelt* [1975]):

$$\begin{pmatrix} a \\ b \end{pmatrix} = \left[ \begin{pmatrix} \mathbf{X} & \mathbf{Y} \end{pmatrix}^* \begin{pmatrix} \mathbf{X} & \mathbf{Y} \end{pmatrix} \right]^{-1} \begin{pmatrix} \mathbf{X} & \mathbf{Y} \end{pmatrix}^* \mathbf{Z} \tag{4.2}$$

($\mathbf{X} = (X_1, \ldots, X_N)$, $\mathbf{Y}$,$\mathbf{Z}$ accordingly. $^*$ denotes the complex conjugate transpose). Here, a program developed by *Egbert and Booker* [1986] has been used, which incorporates a robust scheme that iteratively down-weights data that do not fit the model of Gaussian distributed variables (the *regression M-estimator*, adopted from *Huber* [1981]). The *remote reference* technique (*Gamble et al.* [1979]) additionally accounts for noise in the magnetic components, replacing all righthanded cross-power fields with the horizontal magnetic field of a remote station that recorded simultaneously. For our data, this technique has been helpful for the period range between 10 s and 100 s, where instrumental noise of the deployed fluxgate magnetometers adulterates pure local analysis results (see appendix D).

### Multivariate processing

As explained in sections 2.1 & 2.2, the task of processing array data is basically to find the best estimate of the response space $\mathcal{R}$, i.e. the vector space spanned by the all-component response vectors of the possible linear independent source field potentials (eqs. 2.10 & 2.11). For this multivariate processing, the program written by *Egbert* [1997] has been used.
As in section 2.2, we again postulate that the response space is two-dimensional. Let then

$$\mathbf{X}_i = \mathbf{b}_i + \mathbf{e}_i = \mathbf{U}\boldsymbol{\alpha}_i + \mathbf{e}_i \tag{4.3}$$

be the measured (m×1) data vector containing all array components for the $i$th time segment, or, in a statistical notation, the $i$th realization. Having again $N$ independent identically distributed realizations of the experiment, we can build the matrix of all stacked inter-component cross-spectra, the spectral density matrix (SDM)

$$\mathbf{S} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{X}_k \mathbf{X}_k^* = \frac{1}{N} \left(\mathbf{U}\boldsymbol{\alpha} + \mathbf{e}\right) \left(\mathbf{U}\boldsymbol{\alpha} + \mathbf{e}\right)^* \tag{4.4}$$

$[\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2),\ \boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N),\ \mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_N)]$. Assuming uncorrelated noise of equal variance in all components $(E(\mathbf{e}\mathbf{e}^*) = \sigma^2\mathbf{I}_N)$, the expected value of the SDM gets:

$$E(\mathbf{S}) = \frac{1}{N}\left(\mathbf{U}\boldsymbol{\alpha}\boldsymbol{\alpha}^*\mathbf{U}^*\right) + \sigma^2\mathbf{I} \tag{4.5}$$

Since $\mathbf{S}$ is hermitian, we can formulate the eigenstate problem:

$$E(\mathbf{S})\mathbf{v}_i = \gamma_i\mathbf{v}_i \tag{4.6}$$

*Egbert and Booker* [1989] showed that the eigenvectors $\mathbf{v}_{1,2}$ of the two largest eigenvalues $\gamma_{1,2}$ of this equation span the response space $\mathcal{R} = \ll \mathbf{v}_1, \mathbf{v}_2 \gg = \ll \mathbf{u}_1, \mathbf{u}_2 \gg$, so that the eigenvectors of the two largest eigenvalues of $\mathbf{S}$ give an estimate on the response space. For $\sigma^2 = 0$ (i.e. the data is pure signal) there are just two non-zero eigenvalues.
The assumption of equal noise in all components is generally not justified and in principle problematic, since the components have different units. If more general noise covariance matrices shall be allowed, it is impossible to treat the problem without explicit consideration resp. determination of the noise covariance. To solve this problem, *Egbert* [1997] formulated the generalized eigenstate problem, which according to *Gleser* [1981] yields an unbiased and maximum likelihood (for Gaussian errors) estimate of the response space, if the noise covariance $\Sigma_\mathbf{N} = E(\mathbf{e}\mathbf{e}^*)$ is *known*.

$$\mathbf{S}\mathbf{v} = \gamma\Sigma_N\mathbf{v} \tag{4.7}$$

Without any a priori knowledge, noise which is correlated over the whole array cannot be separated from the signal. Data analysis thus focuses on the careful separation of coherent and incoherent parts of the data. For this purpose, *Egbert* [1997] developed a sophisticated processing scheme, which he called the robust multivariate errors-in-variables (RMEV) estimator. The algorithm iteratively determines the incoherent noise variances, cleans data from outliers (emulating the regression M-estimator from *Egbert and Booker* [1986]) and estimates the coherence dimension of the data. If finally only two eigenvalues of the normalized SDM $\mathbf{S}' = \Sigma_N^{-1/2}\mathbf{S}\Sigma_N^{-1/2}$ of the cleaned data are significantly greater than one, i.e. above noise level, the corresponding normalized eigenvectors ($\mathbf{v} = \Sigma_N^{-1/2}\mathbf{v}'$, with $\mathbf{v}'$ = eigenvalue of $\mathbf{S}'$) can be regarded as the earth's response to approximately uniform sources. Since we defined the signal to be due to pure uniform source excitation, the response to eventual inhomogeneous parts of the source field is considered as coherent noise in this context.

## 4.2 Application to the data sets

No manual data selection has been performed before applying the described processing techniques, relying completely on the robustness of the algorithms. To find a compromise between a high amount of local data samples (time), a high number of components, and a still moderate consumption of computer resources, the simultaneous array data were grouped to sub-arrays of preferably five or six stations for the common processing. Additionally, it was aimed at arranging the arrays in a way that they can all be linked together with as few overlaps, i.e. common stations, as possible (see below). The station runtime plots of the five campaigns in appendix C show the experimental realities, and figure 4.1 illustrates how (far) the goals for the station groupings could be achieved with the given data sets.

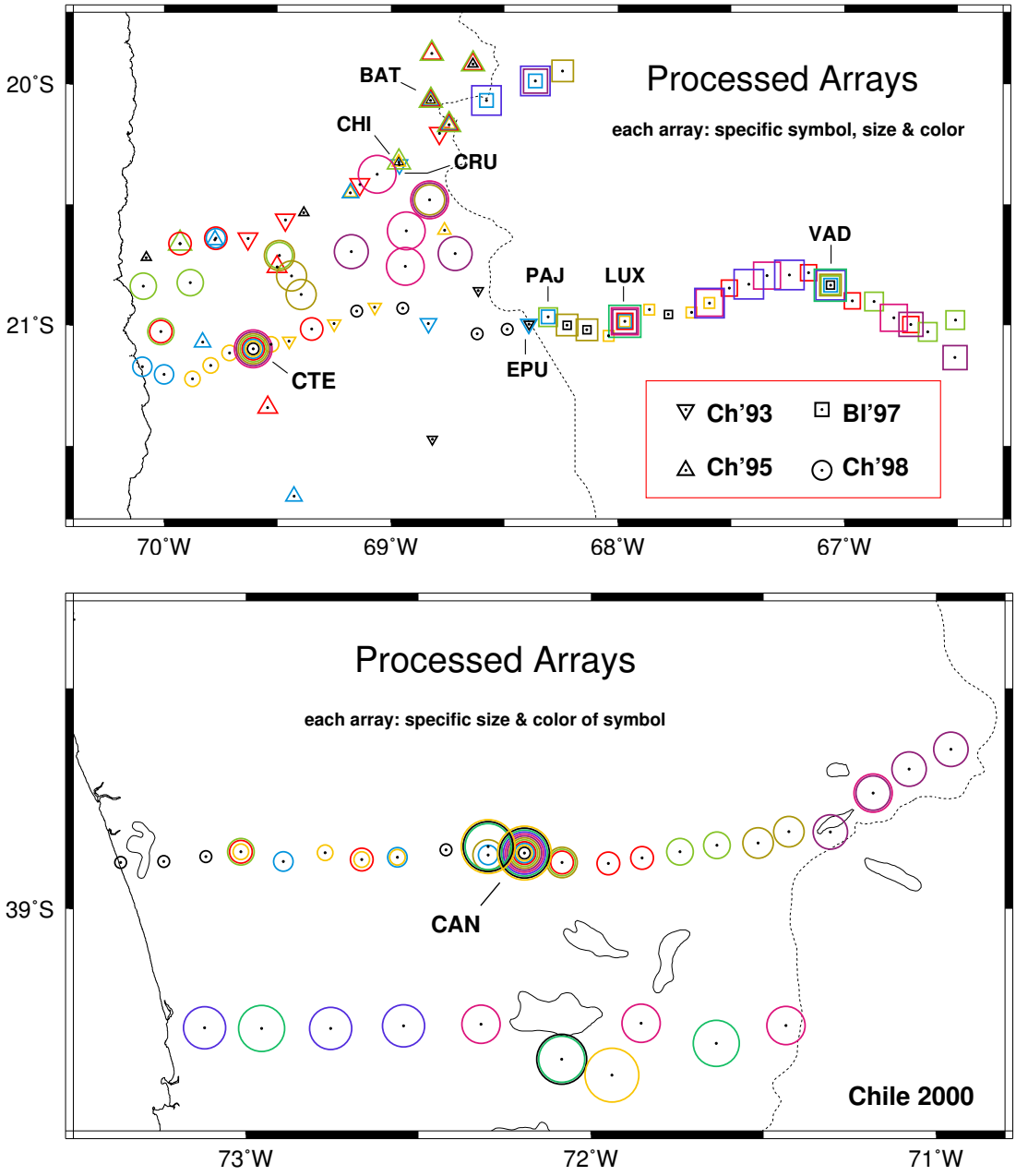


*Figure 4.1: Map showing the processed arrays from the southern Central (∼21°S) and the Southern Andes (∼39°S), each consisting of 3 to 6 stations, mostly 5. When grouping the stations to arrays, it was aspired to find a solution where as few joint arrays as possible cover the whole array with preferably few intersections, i.e. common stations, still using the most part of the data set (see appendix C for station runtimes).*

The 1993 north Chile campaign (14 sites) was the first deployment of the first five RAP data loggers. Accordingly, the small number of equipment together with traditional experimental teething troubles lowered the grade of simultaneity of the corresponding array data, so that only a few, mostly unjoint arrays could be processed successfully. For the north Chile campaigns from 1995 (16 sites, 12 stations) and 1998 (26 sites, 15 stations) and for the south

Chile campaign from 2000, it was possible to group the stations in a way that for both data sets, one distinct station is a member of all processed arrays (BAT & CTE and CAN, resp.). The data from Bolivia 1997 (24 sites, 15 stations) were separated into two groups of arrays for the processing, each of them with one joint station (LUX & VAD, one array includes both sites).

If the first few eigenvalues of the finally calculated normalized spectral density matrices from individual arrays are plotted as a function of period, it is seen that correlated noise is observed in all array data. Furthermore, the structure and level of the respective curves varies significantly from array to array. As a compact illustration, the five dominant eigenvalues from $\mathbf{S}'$ (see above), averaged over all arrays of the respective study area and plotted over period give a good impression of the internal structure of the data (figure 4.2). For the arrays from $\sim 21^\circ$ S, over a wide period range the first two eigenvalues exceed the following ones at least by a factor of 10. Along $\sim 39^\circ$ S at short periods, the signal to noise ratios of the third eigenvalues are less then a factor 3 ($\Delta\text{SNR} \approx 5$) below the second ones, what might be due to cultural noise in this densely populated area. The noticeable increase of the third eigenvalue could be provoked by a violation of the uniform source assumption and be related to a 'normal' vertical magnetic field, as observed by *Egbert and Booker* [1989] in the EMSLAB data set. The general increase of the SNR for the higher eigenvalues towards longer periods might be caused by the decrease of the number of data samples, and thus be a statistical problem (*Egbert, pers. comm.*). Altogether, a calculation of local and inter-station transfer functions from the eigenvectors of the first two eigenvalues should yield at least reasonable results.

As pointed out by *Egbert and Booker* [1989] and *Egbert* [1997], the first two eigenvectors can not generally be regarded as the response vectors to the external source fields of the two polarizations — they just span the same space, if all assumptions on signal and noise are met by the data. However, a graphical illustration of these vectors may well tell us something

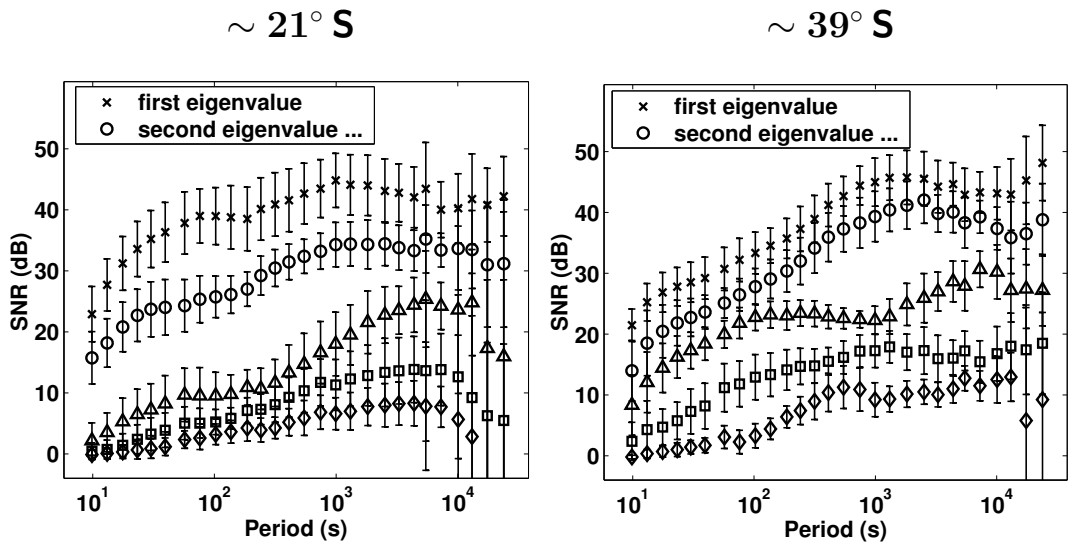


*Figure 4.2: Averages and standard deviations of the five dominant eigenvalues from $\mathbf{S}'$ (see text) of the processed sub-arrays as a function of period.*

about the physical setting they preferably represent. Let $\mathbf{v}^i_{1hor}(\omega)$ be a (2×1) sub-vector of the first eigenvector $\mathbf{v}_1(\omega)$, containing its horizontal magnetic components of site $i$. Then we can easily construct ellipses describing the temporal variation of the horizontal magnetic field for this frequency by plotting:

$$\tilde{\mathbf{v}}^i_{1hor}(\omega, t) = \Re\left[\mathbf{v}^i_{1hor}(\omega) e^{i\omega t}\right] \tag{4.8}$$

(see also *Eggers* [1982]). For joint illustration with the corresponding electrical ellipses, the problem of different units can be treated by a simple conversion according to eq. 2.21, multiplying the electrical field by a factor $\sqrt{\mu_0/\omega\rho}$ (times eventual factors resulting from the mostly preferred units $[nT]$ and $[\mu V/m]$), implicitly assuming a homogenous earth with resistivity $\rho$. Figure 4.3 shows such an illustration for the first two eigenvectors from an array from the north Chile 1998 campaign, comprising five stations located in the Coastal Cordillera and the Longitudinal Valley. We can see that obviously the first eigenvalue bears mainly information of the polarization with the magnetic field parallel to the coastline (TM-mode), whereas the the second eigenvalue refers more to the TE-mode (or at least 'transversal magnetic'-mode). This is observed within *all* arrays from ∼21°S (just in some Bolivian arrays, the first two eigenvalues have about the same level) as well as from ∼39°S, and might be caused by the enormous electrical fields observed due to the ocean-effect in the TM-mode (see section 5.1). For both eigenvectors, the electrical ellipses are not perpendicular to the magnetic ones, which clearly hints at three dimensionality. Yet, dimensionality analysis will be performed purely by analyzing transfer functions.

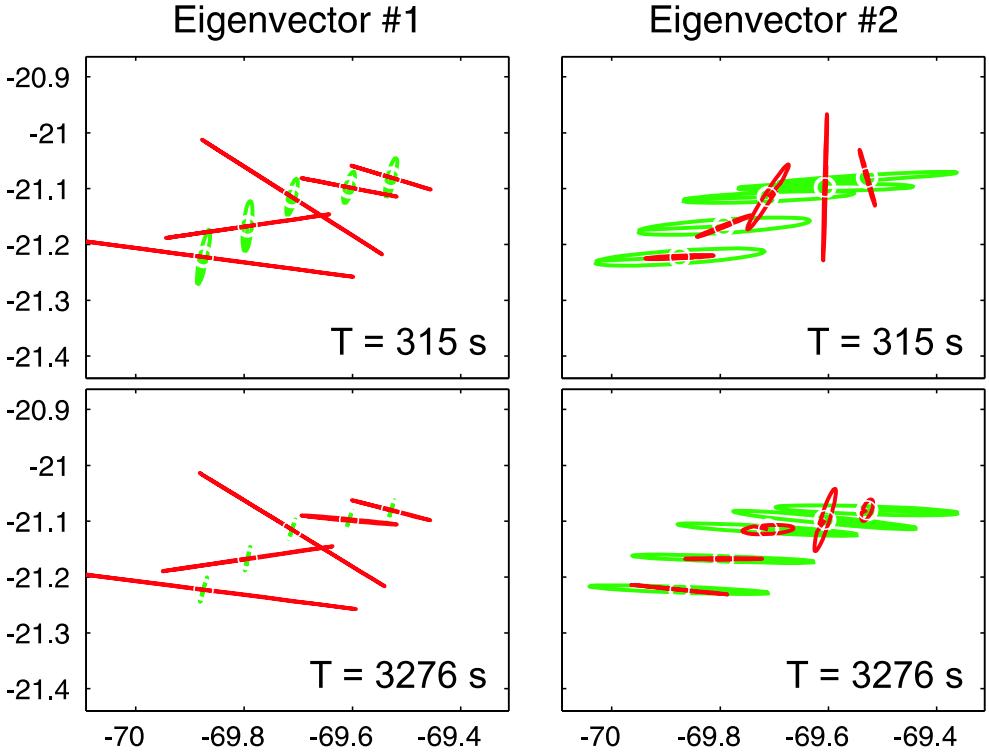


*Figure 4.3: Horizontal components of the eigenvectors from the two dominant eigenvalues from the normalized SDM of an array from northern Chile, including sites LAY, GLO, CDL, CTE and PDT (from west to east).*

## 4.3 Combination of processed arrays

For modelling and/or instructive presentation of the data, it is most convenient to relate all data of the array to a common reference, which in terms of data quality and subsoil conductivity distribution should be chosen with care. To achieve this, a large synthetic array, including preferably all stations, has to be constructed.

If two or more arrays have (at least) one station in common, all components of the arrays can be related to the horizontal magnetic field of this station according to eq. 2.12, and the resulting transfer function matrices can simply be combined. After combination, any other reference field can be chosen. If all small transfer function arrays are spatially overlapping, a plane wave response space for a synthetic array that includes all field stations can be estimated. This could easily be accomplished for the array data from N Chile 1995 & 1998 and southern Chile 2000 by only using one reference station, respectively. As described above, this was not possible for the Bolivian arrays, and the N Chile 1993 data are hard to call array data anyway. Without any spatial overlap, such a procedure is naturally impossible. Fortunately, in this case without considering the possibility of eventual inter-station transfer function analysis, some bad or strange data sites were re-built in following campaigns in order to improve or verify the data, so that the campaigns from the northern Central Andes are at least partially spatially overlapping. To be explicit, The campaigns from 1993 and 1995 nearly overlap at station CRU resp. CHI in the Precordillera, which are just about 1 km apart, and the 1995 campaign intersects with the 1998 field trip at sites BLA and FOR. The biggest handicap for a construction of one final synthetic array is that the Bolivian and Chilean campaigns do not overlap due to the political boundary. For purpose of illustration and/or 'qualitative' modelling, this shortcoming can be treated by regarding the horizontal magnetic fields of the two stations that are nearest to the border — EPU (1993) and PAJ (1997) — as equal.

Though the way to combine the transfer function arrays is in general arbitrary, the temporal and spatial structure of the data set allows just a few reasonable combination schemes, the one that has been applied is illustrated in figure 4.4. Unfortunately, poor data quality of the important overlapping stations CRU and EPU from 1993, together with the missing overlap between Bolivia and Chile impedes a thorough quantitative interpretation of this all-station including synthetic array, and for the inversion of geomagnetic perturbation data, the original arrays and synthetic arrays created with few combination steps and high overlap station data quality will be used (section 7.3). A more sophisticated approach to the combination of arrays, which also considers the information contained in multiple overlaps, is described by *Egbert* [1991]. If such an analysis had been performed, the grouping of field sites to arrays would have been performed differently to take advantage of multiple overlaps. As most data from each campaign are simultaneous, i.e. the transfer function arrays have just to be referenced to one common station to combine them, and overlap between campaigns (if there is any) is not multiple for the most part, this approach does not seem to be really necessary, the more so as data quality is for the most part rather high due to the lack of cultural noise.

After combining the arrays, a normal field has to be defined. In many cases, it may be reasonable to declare the average horizontal magnetic field as normal. This choice is not practical for this data due to the presence of strong conductivity anomalies encountered in

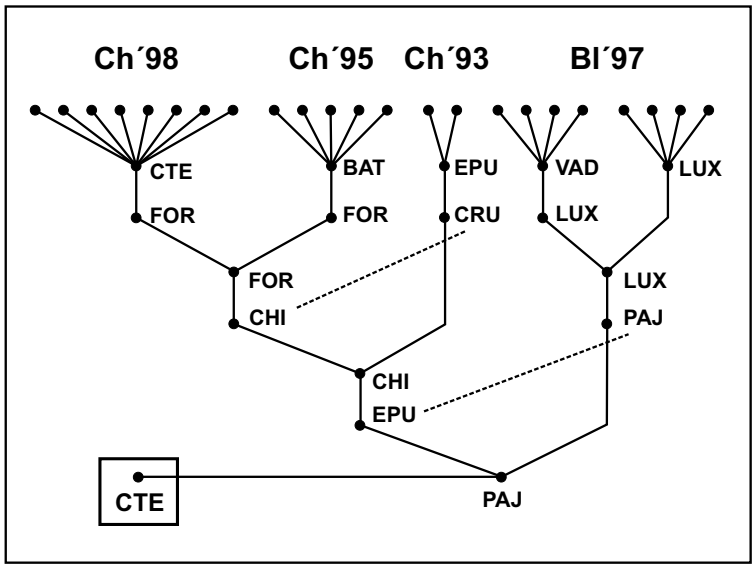


*Figure 4.4: Sketch of combination of the 23 simultaneous sub-arrays (uppermost bullets) from the Central Andes to one final synthetic array. Site names refer to actual reference stations. Dotted lines are drawn between different but closely spaced stations of two disjoint arrays.*

the study area. Therefore, all fields are related to the horizontal magnetic field of one station (CTE), located in the Longitudinal Valley, which operated for 3 months continuously. The resulting transfer function matrix from $\sim$21°S can then be written as:

$$\mathbf{T} \;=\; \mathbf{U}\mathbf{U}^{-1}_{hor,\mathrm{CTE}} \tag{4.9}$$

where $\mathbf{U}$ is now the matrix of the synthetic array, and $\mathbf{U}_{hor,\mathrm{CTE}}$ is a 2×2 sub-matrix of $\mathbf{U}$ with the horizontal magnetic field coefficients from CTE. Transfer functions of the anomalous fields are then easily obtained (cf. eq. 2.16). Note that after this procedure, the transfer functions within the 1998 campaign are exactly the same as before the combination due to the right-side multiplication of an inverse sub-matrix, which underwent the same matrix multiplications as the remaining part from 1998 of the synthetic array.