Variations of southern California seismicity: Empirical evidence and possible physical causes

Jacopo Selva and Warner Marzocchi
Istituto Nazionale di Geofisica e Vulcanologia, Bologna, Italy

Received 22 October 2004; revised 15 July 2005; accepted 23 August 2005; published 10 November 2005.

[1] We investigate southern California seismicity in order to characterize its temporal evolution during the last decades. We analyze the time series composed of the number of events per year and the focal mechanisms of earthquakes since 1933. The results show a statistically significant nonstationarity, with a change that occurred in the 1960s in both time series. The seismicity before the change point is mostly characterized by a strike-slip focal mechanism of San Andreas type; after the 1960s the seismicity appears to show more scattered focal mechanisms and a lower seismicity rate. We provide a possible physical explanation of the significant nonstationarity by modeling the postseismic stress perturbation field induced by the two strongest earthquakes of the last century, the Chile (1960) and Alaska (1964) earthquakes, which both occurred in the 1960s. To first order, the postseismic stress rate seems to be in agreement with the observed changes in seismicity, supporting a causality hypothesis. The model also foretells the future behavior of the trend of southern California seismicity; this forward prediction provides an important opportunity to validate the causal hypothesis of a remote (and long-term) coupling between earthquakes.


1. Introduction

[2] The stationarity of seismic activity is a basic and fundamental assumption of seismic hazard assessments [e.g., Cornell, 1968], and, in general, of models of the spatiotemporal distribution of earthquakes [e.g., Kagan and Jackson, 2000]. From a practical point of view, stationarity means that average and natural variability of the seismicity are constant over time intervals of decades, centuries, and in some cases up to thousands of years as in paleoseismicity studies [Hanks and Schwartz, 1987; Pantosti et al., 1993]. In other words, the seismicity over these time intervals is considered representative of what can happen in the future.

[3] This paradigm implicitly requires that the seismicity of a specific zone may experience significant changes only over larger time intervals (i.e., million of years), on the scale of plate motion processes. On short timescales, the only significant departure from this picture are aftershock sequences; such variations are usually removed from the data set through a declustering technique in order to make the seismic catalog stationary.

[4] Remarkably, in spite of commonly assumed stationarity of seismicity over decades or centuries, some seismic areas show apparent variations over those time ranges. Some studies found that southern California seismicity may have experienced significant changes in long-term activity, even though there is no agreement on the type of variations [cf. Press and Allen, 1995; Jones and Hauksson, 1997; Marzocchi et al., 2003].

[5] Here, we provide some new insights on this topic, by analyzing the southern California seismicity catalog of the last century, which is one of the most detailed and complete catalog of small to moderate magnitude earthquakes. The specific goal is to find and characterize possible nonstationarities from a phenomenological and physical point of view. We analyze the temporal evolution of the seismicity rate, i.e., the number of events per year, and of the focal mechanisms of the earthquakes that occurred in southern California since 1933, looking for statistically significant changes, and quantifying them to provide empirical constraints on physical modeling.

[6] In the second part of the paper we model the changes in seismicity found. In particular, since we are investigating variations over a timescale of decades, the physical model used to describe them has to act over comparable timescales. A possible candidate is the postseismic relaxation of viscoelastic layers beneath the crust [e.g., Piersanti et al., 1997; Pollitz et al., 1998], which has been suggested to be responsible for long-term coupling among earthquakes [e.g., Romanowicz, 1993; Pollitz and Sacks, 1997; Freed and Lin, 2001; Chéry et al., 2001; Casarotti et al., 2001; Marzocchi et al., 2003], and between earthquakes and volcanic eruptions [e.g., Marzocchi, 2002; Marzocchi et al., 2002, 2004].

[7] Here, we model the postseismic effects using a layered, stratified, self-gravitating and viscoelastic Earth model [Piersanti et al., 1995, 1997]. In particular, we estimate the Coulomb failure function (CFF [see Stein et
Due to giant remote earthquakes, which may increase or decrease the tectonic stress loading applied to the faults located in southern California. Note that this approach differs from the one usually followed in studies devoted to stress triggering, where the CFF is used; our approach assumes that the overall rate of seismicity in southern California may be mostly perturbed by the CFF rate evolution rather than its value. We discuss in depth this point later on.

Finally, we anticipate that a relevant aspect of the model is the possibility to validate it in forward analysis since the future evolution of seismicity in southern California may be forecast.

2. Searching for Significant Changes in Seismicity

The seismic database used is the “small box” of the Regional Earthquake Likelihood Models (RELM) catalog, named “Combined provisional southern California catalog point sources” (s_cal catalog), available from Y. Kagan’s Web page (http://moho.ess.ucla.edu/~kagan/s_cal.dat). The catalog reports, for each event, the origin time, the hypocenter location, the magnitude, the focal plane, and the probability $P_m$ that the event can be considered a main shock. The catalog is considered complete since 1933 for magnitude $M \geq 4.7$ [Field et al., 1999] in the so-called small box (see Figure 1). All the characteristics of the s_cal catalog and of the small box can be found in the explanation file available in http://moho.ess.ucla.edu/~kagan/relm.txt. The declustering of the catalog is accomplished by removing all events with $P_m < 0.9$. The choice of this threshold is not critical because almost all the earthquakes have $P_m$ close to 0 or 1 (see Figure 2). In this section, we analyze the sequence of main shocks that occurred in the small box from 1933 to 2003, looking for and characterizing statistically significant changes in seismicity.

2.1. Change Point Analysis

The search for possible change points in seismicity is performed by analyzing two time series: 1) the annual rate of seismicity, i.e., the number of events per year, $(R_t, t = 1, \ldots, T)$, where $T$ is the number of years of the catalog), and 2) the sequence $\rho_i$ defined as $\rho_i = |\sin(\phi_i^3)| (i = 1, \ldots, N)$, where $\phi_i^3$ is the rake of the $i$th earthquake, and $N$ is the number of main shocks in the catalog. We use this transformation of the rake angle in order to transform a circularly distributed variable ($\phi_i^3$) in a new variable ($\rho_i$) that can be treated with classical statistical tests; moreover, the use of the absolute value emphasizes the difference between strike-slip earthquakes ($\rho \sim 0$), and dip-slip events ($\rho \sim 1$).

The change point search is a still unsolved problem in statistics. A detailed discussion on this technical issue is given by Mulargia and Tinti [1985]. In practice, a reasonable strategy to find possible change points has been proposed by Mulargia and Tinti [1985], who applied it successfully in many real and synthetic cases, where no assumptions on the type of statistical distributions of the random variables could be made [see also Mulargia et al., 1987]. In brief, assuming a given significance level in discriminating the different regimes, the method determines the change point according to a sequential scanning which, making use of Kolmogorov-Smirnov two-sample statistics, identifies the principal change point. We refer to this method as CPKS. Here, we also use a modification of the method by using the Wilcoxon test (CPW) instead of the Kolmogorov-Smirnov test, because we specifically want to check differences in the central values (i.e., medians). Technical details of CPKS and CPW are reported in Appendix A.

We apply CPW to $R_t$, and CPKS to $\rho_i$, in order to test, respectively, significant changes in the median of the distribution of the number of events, and in the distribution of the focal mechanism of earthquakes. Despite the method can be used to identify also more than one change point,
here we focus our attention only on the most significant one, i.e., the first one identified by the algorithm (see Appendix A). The results are shown in Figure 3. We find a significant decrease (significance level $\alpha < 0.01$, see Appendix A) of $R_t$ after 1959; this change is consistent with independent results of an analogous test performed on the Southern California Seismic Network (SCSN) catalog in a different region of southern California [see Marzocchi et al., 2003].

In our analysis of $\rho_i$, we find a significant change point ($\bar{\alpha} < 0.01$) in 1969; also this change is consistent with previous results obtained by Press and Allen [1995], who analyzed a different data set. These change points are found also using different values of $P_{m}$ ($P_{m} = 0.5$) and a different transformation of the rakes ($\rho_i = \sin(\phi_i)$).

Note that since the change point is found through a multiple application of tests that are not independent (see Appendix A), $\bar{\alpha}$ cannot be considered as the significance level of the whole change point analysis (hereinafter $\alpha_{CP}$). Nevertheless, the significance level of the change point is constrained to be greater than $\bar{\alpha}$ ($\alpha_{CP} \geq \bar{\alpha}$); thus, given a threshold $\alpha_C$ for rejecting the null hypothesis of no change points in the data, the condition $\bar{\alpha} < \alpha_C$ (see Appendix A) is necessary, but not sufficient, for $\alpha_{CP} < \alpha_C$ (that is for the identification of a significant change point).

Here, this issue is dealt by directly estimating the statistical significance of the change point (i.e., the probability that the change point identified can be observed by chance) through a bootstrap procedure. We generate 1000 synthetic time series of $\rho_i$ (and $R_t$) by shuffling randomly the real sequence; then, we count how many times CPKS (and CPW) gives the most significant change point with a value of $\tilde{\alpha}$ less than the value observed for the real sequence. If we define this number as $m$, the significance level of the change point analysis is $m/1000$. The results of the application of the bootstrap procedure show that the change point of $\rho_i$ in 1969 has a significance level $\alpha_{CP} < 0.01$, while the change point of $R_t$ in 1959 has a significance level $\alpha_{CP} = 0.02$.

Another important issue is to estimate the resolution in time of the change point position. The two sequences show significant change points in 1959 and 1969. Is this difference compatible with the presence of a single change point? In order to give an answer to this question we need to assign a sort of confidence interval to the time location of the change points found. At this purpose, we perform a numerical simulation that provides an indicative measure of the variability of the estimation of the change point location. In particular, we generate 1000 synthetic time series of $R_t$. Each sequence follows a Poisson distribution with different averages before and after a specific change point; the averages are the ones observed in the real sequence $R_t$. The simulation consists of applying CPW to these synthetic times series and collecting the time of the significant change points found. The distribution of the difference between the real position of the change point and the one found by CPW is roughly Gaussian, with $\mu = 0$ year, and $\sigma \approx 5$ years (see Figure 4). Taking into account the $2\sigma$ confidence intervals, we can conclude that the difference of the change point locations found in the two sequences (1959 and 1969) is explained by the limited resolution of the method.

To summarize, our analysis proves the existence of a statistically significant change point in both time series; the limited time resolution of the method locates it in the 1960s.
2.2. Characterization of Seismicity Before and After the Change Point

In order to characterize the statistical change found in southern California seismicity, we divide the seismic data set into two sets, one containing the earthquakes in the period 1933–1959 (C3359), and the other with the earthquakes in the period 1969–2003 (C6903), respectively, before and after the change points found. Since we do not know the exact time of the change point, or if it is gradual or sudden, we cautiously remove the events that occurred in the period between 1959 and 1969 in order to emphasize the differences between the two time periods. Then, we perform a hierarchical cluster analysis (HCA) with Euclidean distances on the sets of focal mechanisms. Basically, the HCA method aims to identify homogeneous groups into a data set, that is, geometrical clusters of data in the parameter space; in other words, a cluster is a group of data that have similar parameter values, and significant differences for data belonging to different clusters. Here, we are interested in understanding whether such kind of clusters exist in both subsets C3359 and C6903, and, in this case, in identifying their peculiar characteristics.

The details of the method are given by Anderberg [1973] and Hartigan [1975] and can be found in Appendix B.

Since HCA does not work on circularly distributed variables, we need to transform the strike ($\phi^{(1)}$), dip ($\phi^{(2)}$), and rake ($\phi^{(3)}$) angles. First, to avoid the coupling between strike and dip angles given by the Aki convention [Aki and Richards, 1980], we apply a 1-1 transformation which make them independent [see, i.e., Selva and Marzocchi, 2004]:

$$S_1 = \phi^{(1)} \quad S_2 = \phi^{(2)} \quad (0 \leq \phi^{(1)} < 180)$$

$$S_1 = \phi^{(1)} - 180 \quad S_2 = 180 - \phi^{(2)} \quad (180 \leq \phi^{(1)} < 360)$$

Then we define the new variables $\theta^{(1)} = \cos(S_1)$, $\theta^{(2)} = \cos(S_2)$, and $\theta^{(3)} = \sin(\phi^{(3)})$ that are not circular.

The centroid of each cluster, i.e., the average focal mechanism, with $\geq$5% of events in the time interval is reported in Table 1. In C3359, two clusters are observed. Cluster 1 covers 93% of the data and is composed of right-lateral events, with almost vertical fault planes and strikes around 139°. The average mechanism is compatible with classical seismicity of San Andreas type [Jones, 1988; Press and Allen, 1995]. Note that here with San Andreas type faults (SAT) we mean all the faults that are almost aligned to the San Andreas fault and have a clear right-lateral mechanism; in other words, SAT includes all the right-lateral faults in southern California (not only the San Andreas) that accommodate the shear tectonic stress. Cluster 2 contains about 7% of the data and it is composed by left-lateral strike-slip events, striking almost perpendicularly to SAT and can be interpreted as its conjugate mechanism. In summary, the clusters identified in C3359 are all linked to SAT system.

The HCA applied to C6903 shows that a significant part of the earthquakes (about 22% of the events in C6903) belongs to clusters that have different characteristics from SAT seismicity. In particular, while cluster 1 contains events with focal mechanisms compatible with the SAT seismicity, clusters 2 and 3 are composed by dip-slip events, occurring on not vertical faults. Note that these two clusters are not observed in the preceding period. These events are the “not SAT” earthquakes (NSAT) which form a pattern in the period 1969–2003 that significantly differs from the preceding seismicity. Our results also confirm the findings of Press and Allen [1995], who, through a pattern recognition method applied to a different data set, found that after the 1960s, many earthquakes began to occur on structures not activated in the previous period.

Table 1. HCA Results Over Subsets C3359 and C6903

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Percentage of Events</th>
<th>Average $S_1$</th>
<th>Average $S_2$</th>
<th>Average $\phi^{(3)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3359</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 1</td>
<td>93</td>
<td>139 ± 12</td>
<td>92 ± 13</td>
<td>174 ± 26</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>7</td>
<td>56 ± 28</td>
<td>76 ± 10</td>
<td>35 ± 46</td>
</tr>
<tr>
<td>C6903</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cluster 1</td>
<td>73</td>
<td>144 ± 24</td>
<td>94 ± 17</td>
<td>-178 ± 27</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>13</td>
<td>121 ± 29</td>
<td>52 ± 13</td>
<td>112 ± 17</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>9</td>
<td>110 ± 22</td>
<td>139 ± 7</td>
<td>108 ± 31</td>
</tr>
</tbody>
</table>

3. Modeling Changes in Seismicity

The results of the analyses reported above indicate a long-term variation, over a timescale of decades. Conse-
coupling among earthquakes because it can increase or decrease the tectonic stress loading rate [see Marzocchi et al., 2003], that is ultimately strictly related to the seismic activity. In other words, we assume that temporal variations in the seismicity rate of a specific type of earthquakes (for instance, SAT) is directly related to the temporal evolution of $\Gamma$.

[25] The use of $\Gamma$ instead of $\Delta CFF$ also accounts for the temporal evolution of stress loading of faults in a proper way. For the sake of example, let us consider a hypothetical case, in which the coseismic (elastic) stress change is positive, while the postseismic (viscoelastic) effect has a negative temporal trend that unloads the fault (i.e., see Figure 5, case 1). In this case, $\Delta CFF$ shows positive values for any time $t > 0$; that is, earthquakes occurring at $t > 0$ are always “promoted” by the perturbation field. Instead, if we consider $\Gamma$, we have a positive value only for the first year after the event (note that the derivative of CFF is infinite at the time of the event, while $\Gamma$ is finite because it is a 1-year filtered derivative) that characterizes the coseismic effects [see Dieterich, 1994], while for longer times the postseismic stress rate $\Gamma$ is negative, i.e., discourages other earthquakes, accounting for the fact that the perturbation works against tectonic loading. Similar discussions can be done for other cases (e.g., see Figure 5, cases 2 and 3), where the relative effects of coseismic and postseismic fields are different.

[26] Here we check the plausibility of this causality hypothesis by calculating the stress perturbations due to the Chile (1960) and Alaska (1964) earthquakes on the fault systems of southern California activated during the past decades. These giant earthquakes are the biggest of the last century, and they occurred at the time of the change point found. Remarkably, southern California has some features that make it suitable to test remote seismic interactions. There is a seismic catalog complete since 1933 for small-moderate magnitudes that is probably the best such catalog in the world. The region is characterized by a high seismic rate, leading to a sufficient number of data to check possible variations. Finally, there are no very big earthquakes (i.e., with $M \geq 8.0$) inside the region that can blur possible effects of remote earthquakes; as a matter of fact, the postseismic variations of smaller events (i.e., Kern County) may produce only localized effects, and they cannot produce a coherent change in seismicity in the whole region as giant remote earthquakes could do.

### 3.2. Chile 1960 and Alaska 1964 Stress Perturbations

[32] The model used to estimate the stress perturbation consists of a spherical, stratified, self-gravitating, and viscoelastic Earth model [e.g., Piersanti et al., 1995, 1997]. The parameters of the model are reported in Table 2. Figure 6 shows the variations of $\Gamma$ induced by the Chile 1960 and Alaska 1964 earthquakes on the characteristic faults defined through HCA and reported in Table 1. Both x and y axes are reported in arbitrary units. The reason is that both values, i.e., time and size of the perturbation, are strongly dependent on the chosen value of viscosity of the asthenosphere, while the general trend is not; in fact, the behavior versus time of $\Gamma$ is similar in shape for all possible values of viscosity. The temporal behavior of $\Gamma$ can be described by four consecutive

### Table 2. Earth Model Parameters

<table>
<thead>
<tr>
<th>Model Parameters</th>
<th>Set Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core radius</td>
<td>3471 km</td>
</tr>
<tr>
<td>Mantle thickness</td>
<td>2620 km</td>
</tr>
<tr>
<td>Mantle Maxwell viscosity</td>
<td>$10^{14}$ Pa s</td>
</tr>
<tr>
<td>Asthenosphere thickness</td>
<td>200 km</td>
</tr>
<tr>
<td>Asthenosphere Maxwell viscosity</td>
<td>$10^{16}$ to $10^{19}$ Pa s</td>
</tr>
<tr>
<td>Lithosphere thickness</td>
<td>80 km</td>
</tr>
</tbody>
</table>
phases, highlighted in Figure 6 (bottom): (1) $G$ promotes NSAT events and discourages SAT events; (2) $G$ produces minor effects on NSAT events and promotes SAT events; (3) $G$ promotes NSAT events and discourages SAT events; and (4) $G$ promotes both NSAT and SAT events. Note that the position of the boundaries of the phases is only indicative, because the transition between phases is gradual.

The first effect forecast by the model is the decrease of the number of events along SAT and an increase of NSAT events (phase 1). This behavior is observed in the data, where after 1959 the number of events decreases significantly, and NSAT events increase their relative occurrence. Figure 7 reports $G_N$ (normal component), $G_S$ (shear component), and $G$ versus time with an asthenospheric viscosity $h = 5 \times 10^{18}$ Pa s, which is a mean value given by previous works [e.g., Piersanti et al., 1995, 1997]. In Figure 7 we can see that both SAT and NSAT faults are locked by the normal stress perturbation induced by Chile (1960) and Alaska (1964) earthquakes; in practice, the earthquakes in southern California were discouraged on average, at least until the end of phase 1, and the consequence of this seems to be the observed decrease of the overall seismicity rate. In the same period (phase 1), NSAT faults undergo greater shear variations than SAT; this may explain the increase of NSAT events observed after the 1960s. It is worth remarking that $G$ values are rather small; this issue will be discussed in depth later on.

With the assumed viscosity of $5 \times 10^{18}$ Pa s, the model sets the transition between phases 1 and 2 around the 1980s–1990s. Our model (see Figure 7, bottom) suggest that this transition is gradual, so that it may become more easily recognizable in the time series when the maximum effect of phase 2 is reached (2010–2020). In a forward perspective, the model forecasts, in absence of other significant perturbations, that the earthquakes of the next few decades should be mostly of SAT (phase 2), like before the 1960s.

Remarkably, this extrapolation of the model is the base of a forward test that may validate the hypothesis of causal relationship between long-term stress changes and nonstationarities in the southern California seismic catalog. The forward test can be performed for both $R_t$ and $\rho_t$ series. Specifically, we can test two null hypotheses. The first null hypothesis consists of equal medians of $R_t$ for the time period 1969–1989 (C6989) and 2000–2020 (C0020), and it can be tested through a Wilcoxon test. Indeed, our model suggests that the median of C0020 should be significantly larger than the median of C6989. The second null hypothesis is of equal distributions of $\rho_t$ in C6989 and C0020, and it can be evaluated through a two-sample Kolmogorov-Smirnov test. Our model forecasts that a significant difference should be found, and that, in C0020, the focal mechanisms of the earthquakes have to be mostly of SAT type as for the period 1933–1959 (see Table 1).

4. Discussion and Remarks

The main finding of this paper is that a statistically significant variation of southern California seismicity occurred in the 1960s. This change consists of a decrease of the overall seismicity rate together with an increase of earthquakes different from the San Andreas type faults. We suggest that a possible physical explanation of such a
change is the remote effect of the giant Chile (1960) and Alaska (1964) earthquakes. The stress perturbation field due to these two remote events is consistent with the trend in the southern California seismicity observed since then. The model is predictive, and therefore it implicitly provides a tool to validate it through forward analysis.

[37] In spite of the agreement in the overall trend, other aspects deserve more investigation. The most important is related to the amplitude of the postseismic stress perturbations; neglecting all the possible biases introduced by our calculations, the stress rate induced by Chile (1960) and Alaska (1964) earthquakes \( \gamma \) is 3 orders of magnitude less than the tectonic rate in southern California (about 10 and \( 10^4 \) Pa/yr, respectively); in other words, it can be considered “small” in relative and absolute sense.

[38] Regarding the “small” value in a relative sense, we argue that it may be misleading to compare \( \gamma \) directly with the tectonic rate. First, for a meaningful comparison we need to project the tectonic rate on the seismogenic structures; this operation may reduce the stress variation of 1 order of magnitude or more [cf. King and Cocco, 2000]. Second, it is certainly more useful to compare the amplitude of \( \gamma \) with other processes that can perturb the system over a comparable time interval, rather than to the tectonic rate directly. Under this perspective, it is worth noting the stability of the tectonic motions measured over time intervals that span 5 orders of magnitude (few to millions of years [e.g., Sella et al., 2002; DeMets et al., 1994]). This may be an important evidence of the extreme stability of the tectonic loading; in this case, the tectonic rate has very low natural fluctuations (at least over time intervals of decades), and therefore it may be significantly perturbed also by apparently small postseismic stress rates. Note that the same point is valid also for static stress changes; in this case, it has been proposed that perturbations as large as tenth of bars may reasonably promote earthquakes [e.g., Reasenberg and Simpson, 1992], also at depth where the lithostatic pressure is also 4 orders of magnitude greater than such a proposed threshold.

[39] Another relevant aspect involves the dimension of the area postseismically perturbed by giant earthquakes; these areas are usually much larger than the one involved by perturbations induced by smaller local earthquakes; in other terms, local earthquakes may yield even higher local stress variations but have a smaller average effect over a large area.

[40] A discussion about the “small” value in an absolute sense implicitly assumes the existence of a stress threshold needed to trigger an earthquake, whose even the existence requires further validation [e.g., Rydelek and Sacks, 1999; Ziv and Rubin, 2000]. Others have considered earthquake nucleation to be part of a critical system and thus highly sensitive to very small perturbations [e.g., Turcotte, 1997]. Under this perspective, we suggest that the only relevant aspect of the stress coupling is to quantify the change in probability of occurrence of earthquakes due to the stress perturbation induced by a (remote) seismic event [e.g., Stein, 1999; Parson et al., 2000; Marzocchi et al., 2003]. In this case, the concept of stress threshold would lose any physical meaning.

[41] Finally, we emphasize that the model can be applied predictively to forecast the future trend of southern California seismicity. In particular, assuming \( \eta = 5 \times 10^{18} \) Pa s, the earthquakes of the next decades should be predominantly of “San Andreas type,” and the rate should increase at a value comparable to the one observed before the 1960s. This forward prediction gives an important opportunity to validate the causal hypothesis of remote (and long-term) coupling between earthquakes.

Appendix A: Change Point Hunting Method Based on the Kolmogorov-Smirnlov Test (CPKS) and the Wilcoxon Test (CPW)

[42] The change point hunting methods aim to find one or more statistically significant change points in a sequence of data. Here, we use two methods based on two different nonparametric statistics; the first one is based on the two-sample Kolmogorov-Smirnlov statistics (CPKS), that is a nonparametric test for equal distributions; the second one is based on the Wilcoxon test (CPW) that is a nonparametric test for equal medians. CPKS has been proposed and tested by Mulargia and Tinti [1985] and Mulargia et al. [1987]; CPW is similar to CPKS, but the Wilcoxon test is used instead of the Kolmogorov-Smirnov test. In practice, CPKS looks for the point along the time axis that marks the most significant change in the statistical distribution of the variable considered. CPW looks for the point in the time axis that indicates the most significant change in the median of the distribution of the variable. Technical details on the Kolmogorov-Smirnov and the Wilcoxon tests are given by Gibbons [1971]. Here, we only remark that in both cases the use of nonparametric statistics allows a wide applicability of the methods in many cases characterized by different (and unknown) statistical distributions.

[43] Let us denote the sequence of data to be analyzed as

\[ x_1, x_2, \ldots, x_N \]  

(A1)

Let us now to chose an index \( j \), so that \( 3 \leq j < N - 4 \). Then, we can divide the sequence in two subsequences

\[ S_1 = x_1, x_2, \ldots, x_{j-1} \]  

(A2)

\[ S_2 = x_{j+1}, x_{j+2}, \ldots, x_N \]  

(A3)

then, we compare \( S_1 \) and \( S_2 \) through a Kolmogorov-Smirnov (CPKS) or a Wilcoxon (CPW) test. These procedures test the null hypothesis of identical distributions or medians respectively (two tails tests) and give the statistical significance \( \alpha_c \) of the differences found, i.e., the probability that the differences found between \( S_1 \) and \( S_2 \) can be obtained by chance.

[44] CPW and CPKS consist of repeating this procedure through all the sequence, i.e., for \( j = 3, 5, \ldots, N - 4 \), and picking the index \( j^* \) that gives the lowest \( \alpha_c \) (\( \alpha_c = \min(\alpha_i) \)). The index \( j^* \) is therefore relative to the most significant change point found in the sequence of data. Given a s.l. threshold \( \alpha_c \), this change point is picked only when \( \alpha_c < \alpha_c \). Further change points can be found by using the same
strategy applied to the two subsets, before and after the jth datum.

Appendix B: Hierarchical Cluster Analysis (HCA)

The Hierarchical Cluster Analysis aims to find "clusters" into a data set, i.e., groups of data that show similarities within each group, and differences with others. This task is accomplished by quantifying a measure that represents the degree of similarity between data. In this case, we use the Euclidean distance between data represented in the parameter space, i.e., a space with m dimensions, where m is the number of parameters that characterize each datum.

First, HCA treats each datum as a cluster. Then, the two closest clusters (i.e., the two clusters with the minimum distance) are merged, and the distance between the new clusters is computed, by means of the average-distance-between-clusters method [Anderberg, 1973]; in brief, each cluster is represented by a point in the parameter space that represents the average of the data belonging to the cluster, and the distance between clusters is the Euclidean distance between these average points. This aggregation process continues until only one cluster remains. In this way, a binary cluster tree is defined. Given a fixed number of clusters $N_{cl}$, and the cluster tree, the cluster membership of each observation is finally computed.

The main problem is to define the right number of clusters. At this purpose, we analyze the graph of the within-clusters variance $W$ as a function of the number of clusters $N_{cl}$. $W$ is computed by

$$W(N_{cl}) = \sum_{j=1}^{N_{cl}} \frac{N_{dat,j}}{N_{cl}} (\bar{x}_j - \bar{x})^2$$

where $\bar{x}_j$ is the vector of parameters of the jth datum belonging to the jth cluster, $\bar{x}$ is the average point of the jth cluster, and $N_{dat,j}$ is the number of data in the jth cluster. In other words, $W$ is a direct measure of the variance "explained" by the model with $N_{cl}$ clusters (the so-called within-clusters variance). Remarkably, the plot of $W$ as a function of $N_{cl}$ is very useful to determine the number of clusters into the data set [e.g., Priestley, 1981]. For instance, if we consider a number of clusters smaller than the true value, we may expect that $W$ will be larger than the true variance (associated to the true number of clusters), because any additional cluster omitted from the model would "explain" a further part of the variance of the variable $x$. On the other hand, once the right value of $N_{cl}$ is reached, any further increase in the number of clusters will not significantly reduce $W$. Hence we expect that the plot of $W$ as a function of $N_{cl}$ will decrease at first, and then "level out" at the point where $N_{cl}$ reaches the true number of clusters.

Acknowledgment. The authors thank two anonymous reviewers for helpful comments and suggestions.

References


W. Marzocchi and J. Selva, INGV-Bologna, Via D. Creti 12, I-40128 Bologna, Italy. (marzocchi@bo.ingv.it; selva@bo.ingv.it)