USGS
science for a changing world

Analysis of a Spatial Point Pattern: Examining the Damage to Pavement and Pipes in Santa Clara Valley Resulting from the Loma Prieta Earthquake

By G.A. Phelps

Open File Report 2007–1442

2008

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
Dirk Kempthorne, Secretary

U.S. Geological Survey
Mark D. Myers, Director

U.S. Geological Survey, Reston, Virginia 2007

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment:
World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS

Suggested citation:
Phelps, G.A., 2007, Analysis of a spatial point pattern: examining the damage to pavement and pipes in Santa Clara Valley resulting from the Loma Prieta earthquake: U.S. Geological Survey Open File Report 2007-1442, 49 p.
[http://pubs.usgs.gov/of/2007/1442/].

ABSTRACT

This report describes some simple spatial statistical methods to explore the relationships of scattered points to geologic or other features, represented by points, lines, or areas. It also describes statistical methods to search for linear trends and clustered patterns within the scattered point data. Scattered points are often contained within irregularly shaped study areas, necessitating the use of methods largely unexplored in the point pattern literature. The methods take advantage of the power of modern GIS toolkits to numerically approximate the null hypothesis of randomly located data within an irregular study area. Observed distributions can then be compared with the null distribution of a set of randomly located points. The methods are non-parametric and are applicable to irregularly shaped study areas. Patterns within the point data are examined by comparing the distribution of the orientation of the set of vectors defined by each pair of points within the data with the equivalent distribution for a random set of points within the study area. A simple model is proposed to describe linear or clustered structure within scattered data.

A scattered data set of damage to pavement and pipes, recorded after the 1989 Loma Prieta earthquake, is used as an example to demonstrate the analytical techniques. The damage is found to be preferentially located nearer a set of mapped lineaments than randomly scattered damage, suggesting range-front faulting along the base of the Santa Cruz Mountains is related to both the earthquake damage and the mapped lineaments. The damage also exhibit two non-random patterns: a single cluster of damage centered in the town of Los Gatos, California, and a linear alignment of damage along the range front of the Santa Cruz Mountains, California. The linear alignment of damage is strongest between 45º and 50º northwest. This agrees well with the mean trend of the mapped lineaments, measured as 49º northwest.

INTRODUCTION

Scattered point data are common in the geological sciences; earthquake epicenters, the locations of mineral deposits, and oil plays are examples of geologic data that are represented as point data. Two questions are common with these data: “Are the points related to some other geologic feature?” and “Do the points themselves exhibit a pattern?”

One of the difficulties of statistically analyzing scattered datasets is that the data often occur within irregular study areas. A null hypothesis of randomness, in this case a set of randomly located points, is often the starting point for any analysis. C*omplete spatial randomness* for points is defined as a set of points whose locations are an independent random sample taken from a distribution of equal probability across the study region (Diggle, 2003). This implies that the location of a given point is independent of the location of any other point. In other words, the points do not interact with one another.

A typical method of generating distribution parameters of point data sets satisfying complete spatial randomness within an irregular study area is Monte Carlo simulation, where randomly located points are added to a study area and the properties are averaged over 100 or more simulations (Diggle, 2003). While this method is robust, it often requires programming effort to meet the needs of the particular analysis, requires computing power, and does not necessarily offer any insight into the problem.

Some distributions can also be estimated more directly than by the use of Monte Carlo simulations. Such direct methods commonly require less computing time, less programming effort, and ultimately offer more insight into the problem at hand. As will be seen, for example, distributions based on a null hypothesis of complete spatial randomness can be conveniently estimated in at least two cases by using powerful raster processing tools available in modern GIS software packages.

The first case, the problem of investigating the spatial relationship of scattered point data to other objects, was described by Okabe and Fujii (1984). The second case, the problem of investigating the randomness of a scattered dataset itself, can be derived from principles developed in solving the first case. The following sections are devoted to exploring methods of approaching both problems, using a dataset of damage to pavement and pipe breaks caused by the 1989 Loma Prieta earthquake as the point dataset, and a set of mapped areal photographic lineaments as the objects which may be spatially related to the earthquake damage.

DATA AND SOFTWARE

The 1989 M_w 6.9 Loma Prieta earthquake (U.S. Geological Survey, 2007), the epicenter for which was located roughly 100 kilometers south of San Francisco, California, caused damage to many types of public works, including roads and sidewalks, throughout Santa Clara Valley (figure 1). The location and type of damage were mapped via extensive fieldwork and database compilation (Schmidt and others, 1995). 1427 observations, taken from fieldwork, the records of utility and transportation institutions, and local governments, were recorded for a 663 square kilometer area. Schmidt and others (1995) exhaustively searched roads and parking lots within the study area, measuring contractional damage indicators for ground level damage, pavement breaks and curb breaks. They added to this field data by incorporating damage for sub-surface gas and water line ruptures, data shared by local utility, transportation, and governments. In all, five types of damage were recorded (asphalt, channel lining, concrete, gas line, water line), in addition to the sense of deformation (if any), the freshness of the damage, and whether or not the damage was at or below the ground surface. Damage to structures was not included because it depends upon building construction, materials, and design, and the damage cannot be constrained to the Loma Prieta event. Sidewalks, pavement, and other public infrastructure works are more commonly built to uniform specifications and can therefore be used to detect ground motion in a consistent manner. In this report only the location of the damage is considered in the analyses; analyses based on subsets of the data would be a reasonable next step for further research. For further information discussing the 1989 Loma Prieta earthquake the reader is directed to U.S. Geological Survey Professional Papers 1550-1553.

The (unpublished) digital version of damage recorded in Santa Clara Valley is used as the example scattered data for analysis in this report. While the data were collected by plotting the locations on 1:24,000-scale topographic maps (before the widespread use of GPS), subsequent digitizing and registration to scanned topographic maps indicates a relatively high level of positional accuracy, where data points often plot on the correct side of the street. While not all points were checked, this observation suggests the data points are accurately located to within a few meters.

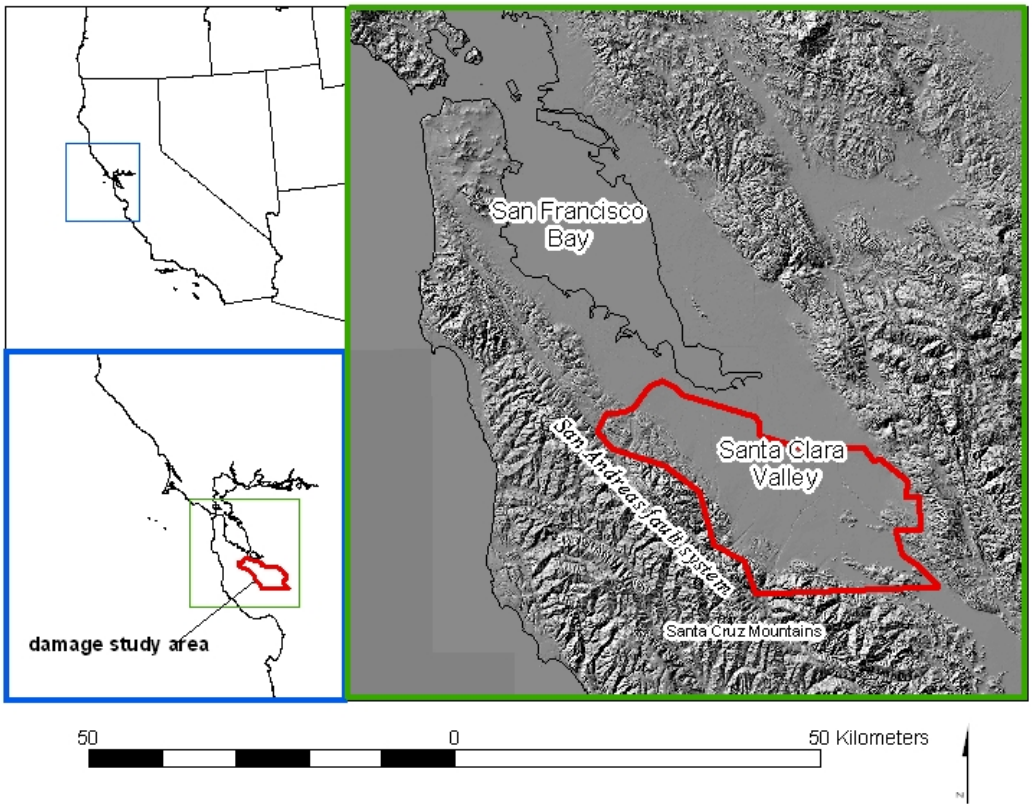


Figure 1. Location of the Santa Clara valley in the San Francisco Bay area, California, USA. Damage used in this report from the Loma Prieta earthquake was recorded within the study area shown in red.

Hitchcock and others (1994) mapped a series of areal photographic lineaments as part of a Quaternary geologic mapping study to investigate seismic activity along the eastern range front of the Santa Cruz Mountains, where the mountains abut Santa Clara Valley. The lineaments are based on several types of geomorphic and photographic features, such as topographic scarps and depressions, stream sinuosity, vegetation lineaments, and tonal changes in the ground surface. The lineament database used in the next section is an unpublished digital rendition of geology from plate 2 of Hitchcock and others (1994).

The GIS software used to test and implement the methods described in this report was the commercial package Arc/Info, and the associated product ArcMap (ESRI™). Arc/Info has built-in tools to generate many of the analytical results used in this report, including density maps, buffers, nearest-neighbor distances, point-in-poly operations, and database merges and queries. Appendix II contains pseudo-code for the analytical operations performed in this report.

The commercial statistical software package Splus (Insightful™) was used for the statistical analysis, including graphical displays, analysis of distributions, and goodness-of-fit tests.

METHODS AND RESULTS

In order to compare the observed pattern of a scattered point dataset to a null hypothesis of a pattern of complete spatial randomness one must be able to describe the null distribution for the study area in question. This is often done in the literature by means of quadrat analysis, where the study area is broken up into squares of equal size (quadrats) and the frequency of the points per quadrat generates a Poisson distribution under complete spatial randomness (Diggle, 2003; Upton and Fingleton, 1985). One is not always fortunate enough to have a rectangular study area, however. The natural world often precludes rectangular study areas by imposing natural boundaries (e.g. rivers and lakes, forests, cliffs and other steep terrain), and existing datasets were not always created to conform to a rectangular study area. If one does not wish to be limited in analyzing scattered datasets, methods must be developed to describe the distributions of complete spatial randomness for irregular study areas.

The question "Are the points in a given scattered dataset related to some geologic phenomenon?" can be addressed by considering the geologic phenomenon as fixed, and the points as realizations of some process that is perhaps influenced by the phenomenon. Points are considered "events" that occur with some probability within the study area. For the null hypothesis of complete spatial randomness, the study area is considered homogenous; that is, a point has equal probability of being located anywhere within the study area, independent of features (e.g. topography, soil type, vegetation). Now suppose that this assumption is violated: some geologic phenomenon creates inhomogeneities within the study area or some geologic process causes the points to locate non-randomly within the study area. If the points and the geologic phenomenon are related, one might expect either a positive or negative spatial association between them, in which the points would either have a tendency to be attracted to, or avoid, the phenomenon. In other words, in the end-member cases the points will either be co-located with the phenomenon or as far from it as allowable given space constraints. The strategy, then, is to find a null distribution for the distance from the geologic phenomenon, under the assumption of complete spatial randomness, within the given study area, and compare this with the observed distribution. If the observed distribution of distance from the geologic phenomenon is different from the null distribution, then one can determine if two are positively or negatively correlated.

The question "Are the points themselves randomly located within the study area?" amounts to determining whether or not the scattered point dataset can be distinguished from complete spatial randomness. A new method of examining whether a scattered dataset deviates from complete spatial randomness is described. This method relies on the orientation specified by pairs of points within the dataset.

Complete spatial randomness depends on the sample points being independent of one another. No point may have an effect on the location of another point. Stated differently, if there are two processes influencing the location of points, then the effect from one may influence the statistical results calculated for another, and this would violate the assumption of independence (of course it would be very unusual to have only one process operating in any natural system, but one would like the dominant process to overshadow the rest). Before any analysis begins the dataset should be examined for obvious effects from secondary processes.

Association of damage points with areal photographic lineaments: an example from the Loma Prieta earthquake

"Are the points in a given scattered dataset related to some geologic phenomenon?" is addressed for the example of comparing damage incurred during the 1989 Loma Prieta earthquake in Santa Clara Valley, CA (figure 2),

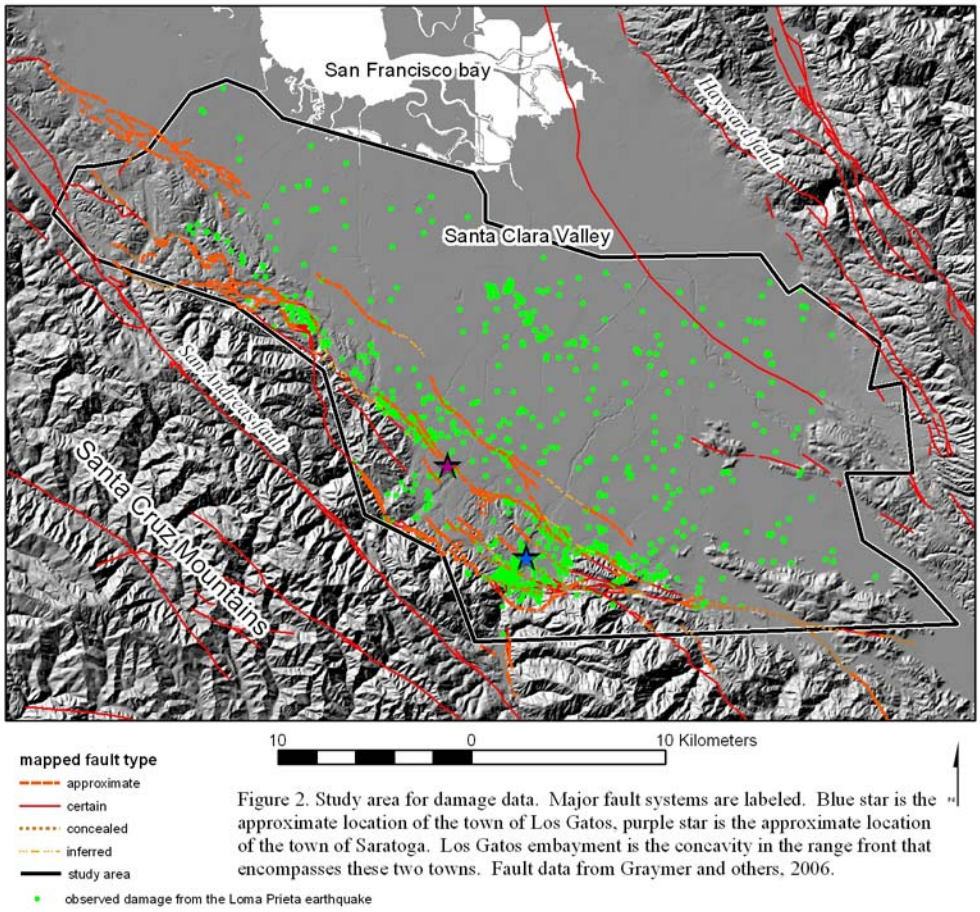


Figure 2. Study area for damage data. Major fault systems are labeled. Blue star is the approximate location of the town of Los Gatos, purple star is the approximate location of the town of Saratoga. Los Gatos embayment is the concavity in the range front that encompasses these two towns. Fault data from Graymer and others, 2006.

to a set of areal photographic lineaments mapped within an irregular study area along the range front of the Santa Cruz Mountains (Hitchcock and others, 1994), near the towns of Los Gatos and Saratoga (figure 3). The study boundary is determined by using the full extent of the available geologic mapping presented by Hitchcock and others (1994), a somewhat irregular strip of mapping along the range front of the Santa Cruz Mountains.

The example demonstrates the method for determining a spatial association between linear features and a scattered point dataset, but the method is easily extendable to point features or area features (Okabe and Fujii, 1984; Okabe and others, 1988). For convenience, references to damage in the rest of this report will mean the damage from Loma Prieta earthquake in Santa Clara Valley as describe by Schmidt and others (1995), and the study area covered by the damage data will be referred to as the "damage study area." Similarly, references to lineaments will mean the lineaments mapped by Hitchcock and others (1994) and the study area encompassing the lineaments will be

referred to as the “lineament study area.” The lineament study area is completely contained within the damage study area (figure 3).

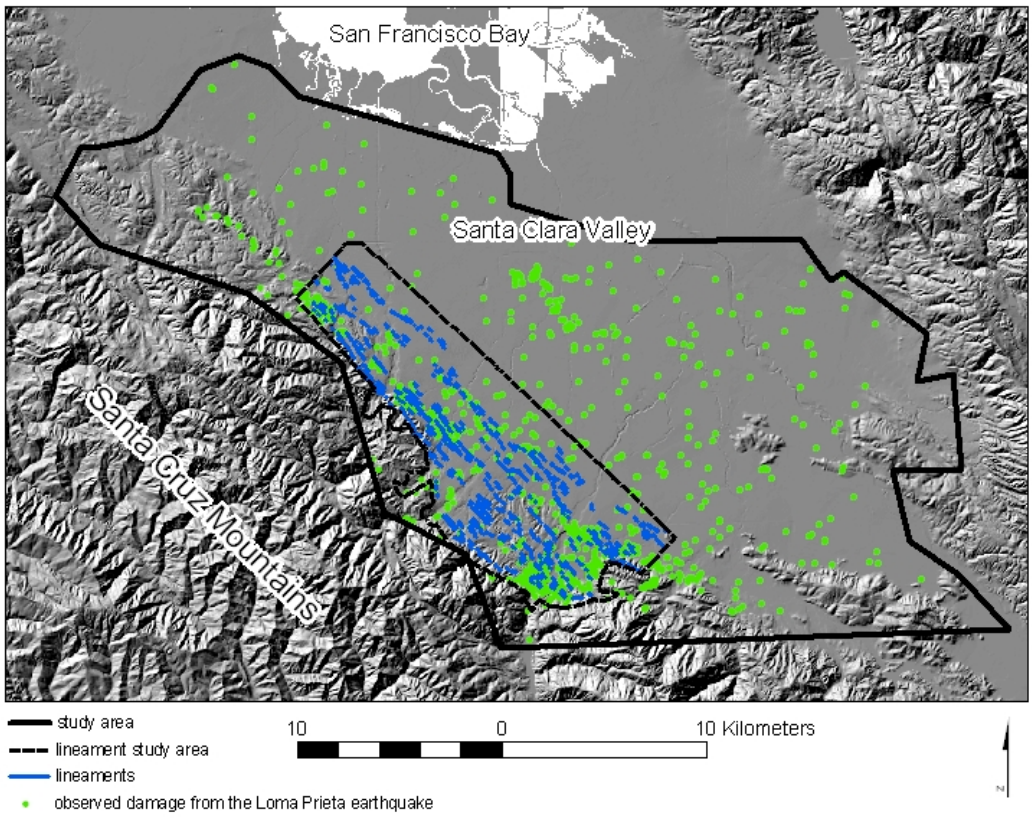


Figure 3. Study area for damage data and the lineament data. The study area for the lineaments is a sub-area of the study area for the damage.

The lineament data have been cited as evidence to support the interpretation and mapping of reverse faults along and outboard of the range front of the Santa Cruz Mountains (Hitchcock and others, 1994; Hitchcock and Kelson, 1999). This interpretation is supported by geophysical evidence of a steep-sided subsurface basin adjacent, and parallel to, the range front (Langenheim and others, 1997). However, while some of the lineaments mapped are based on observed geologic features such as topographic scarps, others are based on vegetation or tonal changes seen in areal photographs. The latter types of evidence are not necessarily indicative of fault activity. A spatial association of the damage with the lineaments would strongly suggest the features that define the lineaments are related fault activity. If the damage is, on average, closer to the lineaments than would be the case for complete spatial randomness, then the lineaments and the damage tend to be co-located and are thus spatially related.

To answer the question of whether or not the damage is associated with the lineaments a model of randomness must be constructed with which to compare the observed data. Such a model can be developed by using the distance of each damage point to the nearest lineament. If the damage and the lineaments tend to be co-located, the damage ought to occur on or near the lineaments. If they avoid co-locating, the

damage ought to be as far as possible from the lineaments. A spatially random distribution of damage locations would show no association with lineaments throughout the study area.

The distribution of the distance of the damage to the lineaments can be obtained by measuring the distance of each damage point to the nearest lineament (considering only the subset of the damage data that falls within the lineament study area (as defined by Hitchcock and others, 1994)). The problem then becomes how to generate the probability density function of the distance to the lineaments for a random process. Once this is accomplished the two distributions can be compared.

Assume the lineaments are fixed; that is, they are mapped correctly and properly located. This is not an insignificant assumption, but here the mapped features are taken as given. Then, over the lineament study area, one can compare the distributions of the distance of damage points to the nearest lineaments with the distance expected for a randomly distributed set of points.

To obtain a random distribution, one need only consider the probability of a point falling in a particular sub-region of a given study area (Okabe and Fujii, 1984). Since a random point has equal probability of falling at any particular place in the study area, the probability that it will fall in a given sub-region is simply the area of the sub-region divided by the total area (figure 4).

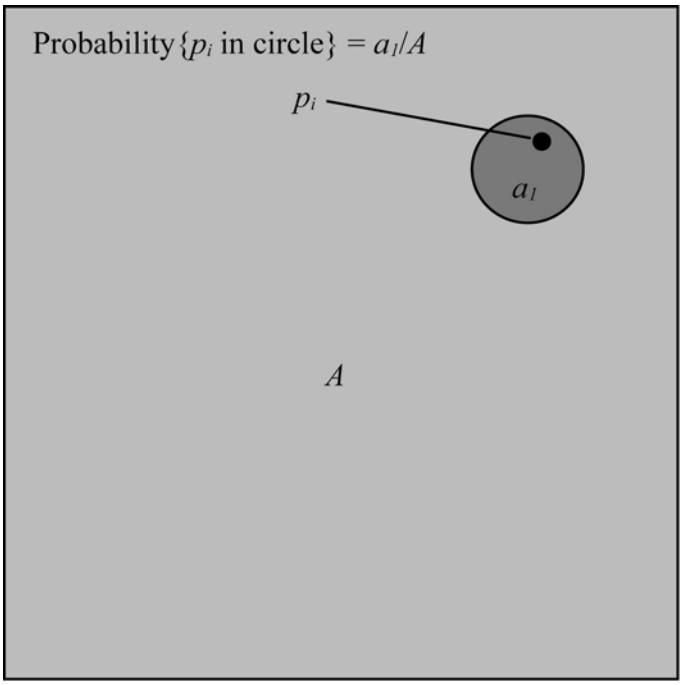


Figure 4. The area of the square is defined to be A, and the area of the circle is defined as a_l. The probability that a randomly located point, p_i, within the square falls within the circle is a_l/A.

Now consider the sub-region of the study area defined as any point greater than x and less than $x+h$ distance from the nearest lineament (figure 5). The probability of a

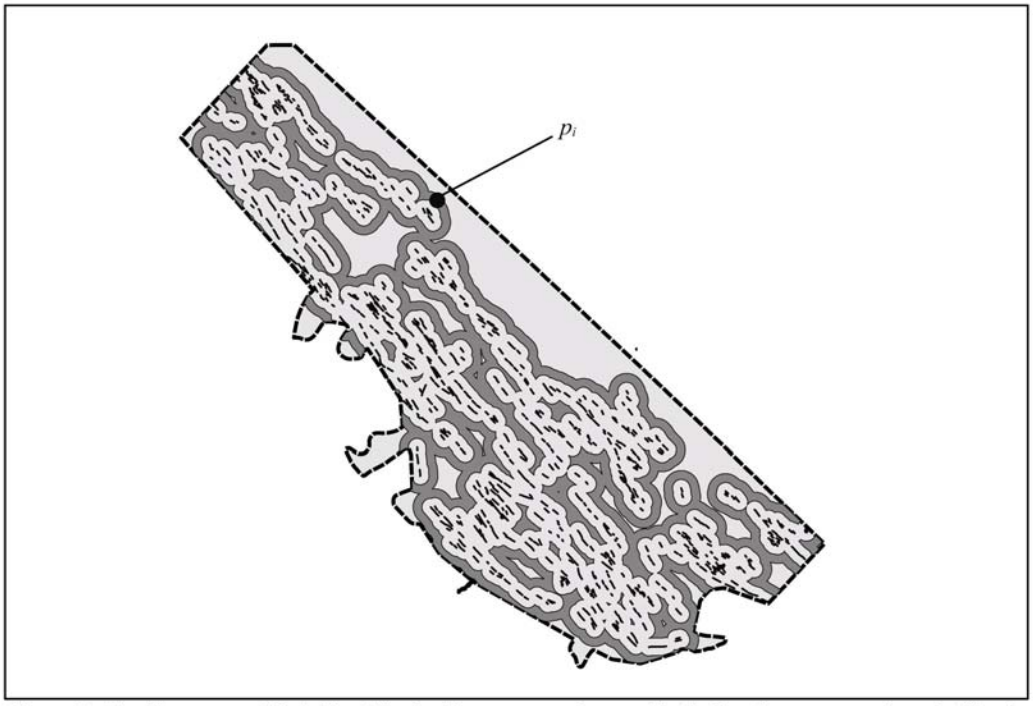


Figure 5. The lineaments (black lines) in the lineament study area (dashed outline, area = A) are buffered, showing the sub-region that is between 250 m and 500 m from the nearest lineament (dark grey, area = a_l). The probability that a randomly located point, p_i, will fall within the buffered region is the area of the buffered region divided by the total area, or $Pr\{250<p_i<500\}=a_l/A$, where the numbers represent the distance to the nearest lineament, in meters.

random point being between x and $x + h$ distance from the nearest lineament in the lineament study area is simply the area defined by the buffer around the lineaments from x to $x + h$ divided by total area. If the probabilities for many buffers (for example, in 200 m increments) are combined for a sequence of distances from zero to the farthest distance in the study area, the result is an approximation of the probability density function for the distance to the nearest lineament for a random set of points.

This distribution can be numerically approximated by rasterizing the lineament study area and keeping track of all cells a given distance from the nearest lineament (figure 6). The ratio of the area of the sub-regions between x and $x + h$ distance from the nearest lineament to the total study area is simply the number of pixels between x and $x + h$ distance from the nearest lineament divided by the total number of pixels in the study area (the units cancel). This numerical approximation of the random distribution can be compared to the observed distribution of the distance of the damage points to the nearest lineament. Since the resulting probability density functions are non-gaussian, a Smirnov test (Rock, 1988) can be used to test whether or not the two distributions are different.

The above method is appropriate provided one assumption is satisfied: that it is assumed the observed process is, or at least could be, homogeneous over the study area. That is, the process is the same at any location in the study area. For example, as a first approximation, damage is assumed to be equally likely regardless of the type of soil that

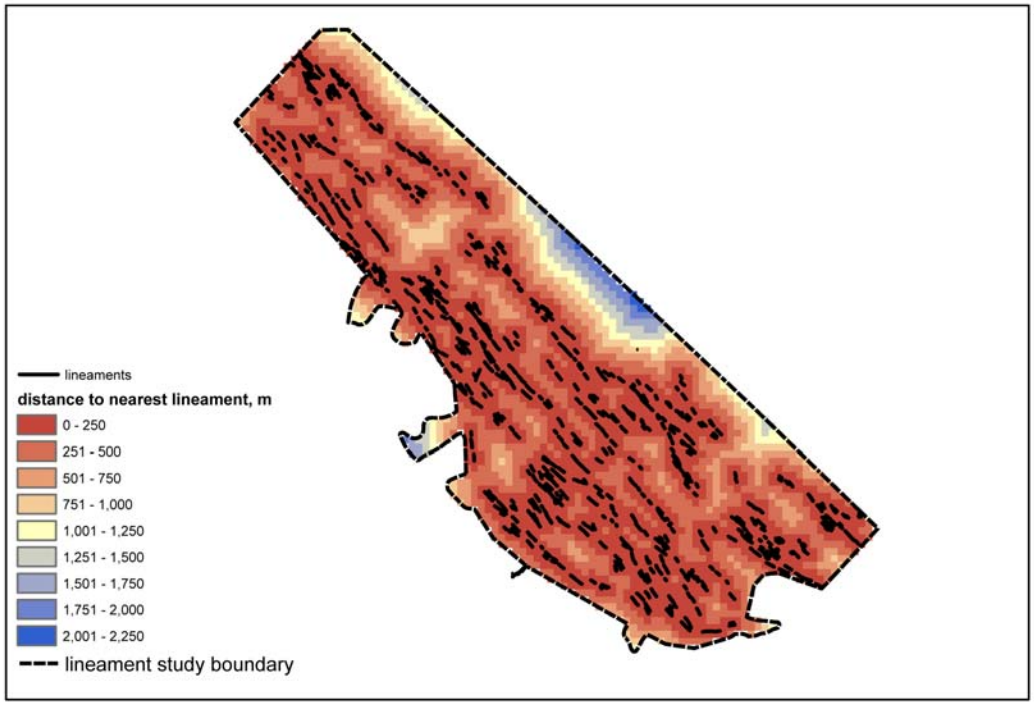


Figure 6. The lineament study area is rasterized (converted to pixels), and each raster cell (pixel) is assigned a value equal to the distance of the center of the pixel to the nearest lineament (in meters). By plotting the distribution of distances, one can approximate the probability density function of distance to the nearest lineament for a randomly located set of points.

is present. If the process is homogeneous (does not depend on soil type), it is then reasonable to compare this observed distribution with a random distribution that meets the same assumption of homogeneity. A study area that includes sub-areas that affect the process differently is called inhomogeneous.

The mapped damage is an example of an inhomogeneous point pattern, because the damage data consists of pavement breaks (these occur exclusively on paved areas i.e.: streets and sidewalks) and pipe breaks (also almost always occurring beneath streets, because utilities tend to follow public streets and rights-of-way). Practically speaking, this damage cannot occur outside of an area covered by a street. The process is therefore inhomogeneous over the both damage and lineament study area. The total sample space is the subset of the study areas that are covered by streets and sidewalks.

Changing the sample space changes the probability of a random point falling a given distance from a line because it changes the total area available to the random point. In order to calculate the probability for the inhomogeneous point process the area of the streets and sidewalks must be found.

Finding this subset of the lineament study area would require coupling extensive data at the county level for street footprints, if such data could be obtained. A scanned 1:100,000 USGS topographic map of the streets for the lineament study area gives a

reasonable approximation of the area covered by streets and sidewalks, as can be seen in figure 7,

Figure 7 Detailed view of the scanned USGS topographic map of streets overlain on the digital orthophoto quadrangle for a portion of the study area. Note how the scanned streets correspond reasonably well to the area of streets on the digital orthophoto both in location and in area. Note also the errors from areas of recent development (lower right), where the streets are missing in the scan. The damage is located near, but not perfectly on, the streets, the result of errors in digitizing.

which compares the scan to the 1995 digital orthophoto quarter-quadrangle for a portion of the area. While certainly not error-free, the scan appears to approximate the area to within 10% to 20%. In the subsequent analyses the scan will be used to generate the subset of the lineament study area for examining the inhomogeneous point process.

Note that the lineaments are assumed to be fixed, and the damage is taken as the process that is either related to these fixed features or not. The question could have been posed conversely; are the lineaments located more closely to the mapped damage than a random process for generating lineaments? In this case the damage points are the more stable features. Based on geomorphologic evidence, the lineaments represent features that span a range of ages (Hitchcock and others, 1994). If they are tectonic features, then one would expect them to span multiple events. In this case the damage represented by a single event would not necessarily be associated with every lineament, but with a subset belonging to one or more fault strands experiencing activity in the Loma Prieta event. Therefore the question is posed such that the particular event is compared with the general tectonic framework.

DECLUSTERING THE DATA

The method described above assumes that the damage is not significantly affected by another process. A density plot[1] of the damage data (figure 8, top) shows the linear concentration of damage along the range front of the Santa Cruz Mountains apparent in figure 2, and the tight clustering of damage near the town of Los Gatos. Previous authors (Hitchcock and Kelson, 1999; Langenheim and others, 1997; Schmidt and others, 1995) have noted the linear nature of the damage, and Schmidt and others (1995) noted the apparent clustering of damage, with 54% of the damage occurring in the Los Gatos 7.5' quadrangle. What is apparent in the density plot, and perhaps less apparent in figure 2, is the magnitude of the clustering relative to the linear concentration of damage. The cluster of data near Los Gatos is by far the dominant signal in the point pattern, with the density of points almost four times larger than the density along the linear concentration of damage. The cluster is at the southwest end of the Los Gatos embayment, a concavity in the range front near the town of Los Gatos (figure 2). In addition to lineaments, several mapped faults are also present within the region of the cluster. Is the cluster of damage due to local faulting, or another process, such as the shape of the embayment, which perhaps concentrated and focused seismic waves? The interpretation of the statistical results depends on the answer to such questions, and the questions are typically not easy to answer. If the clustering in the damage data (near Los Gatos) is due to a process independent of that which is potentially generating the lineaments, then the clustering will bias the resulting statistical analysis.

In order to account for the possible bias of the cluster, it was modeled as a circular anomaly and the effect removed from the data. This was accomplished by examining the characteristic location and shape of the cluster observed in the density plot. The density plot shows the cluster of damage has a locus within the town of Los Gatos. Furthermore, the cluster is a roughly circular phenomenon that appears almost isolated from the rest of the damage. This can be seen in a cross-section of the density plot of figure 8 (top) that transects the cluster (figure 9). A gaussian curve fit to this cross-section models the cluster well to two

[1] Density plots in one dimension are "smoothed" histograms, created by counting the number of data points within a moving window (rather than within a fixed bin as the histogram does), and applying a weighting function to the points so that points at the edge of the window contribute less to the total count within the window. In two dimensions the moving window is in the shape of a circle, typically tapering at the edges using a gaussian or quadratic kernel. See Silverman (1986) for a complete description of density plots.

standard deviations. The model indicates that the cluster has a center

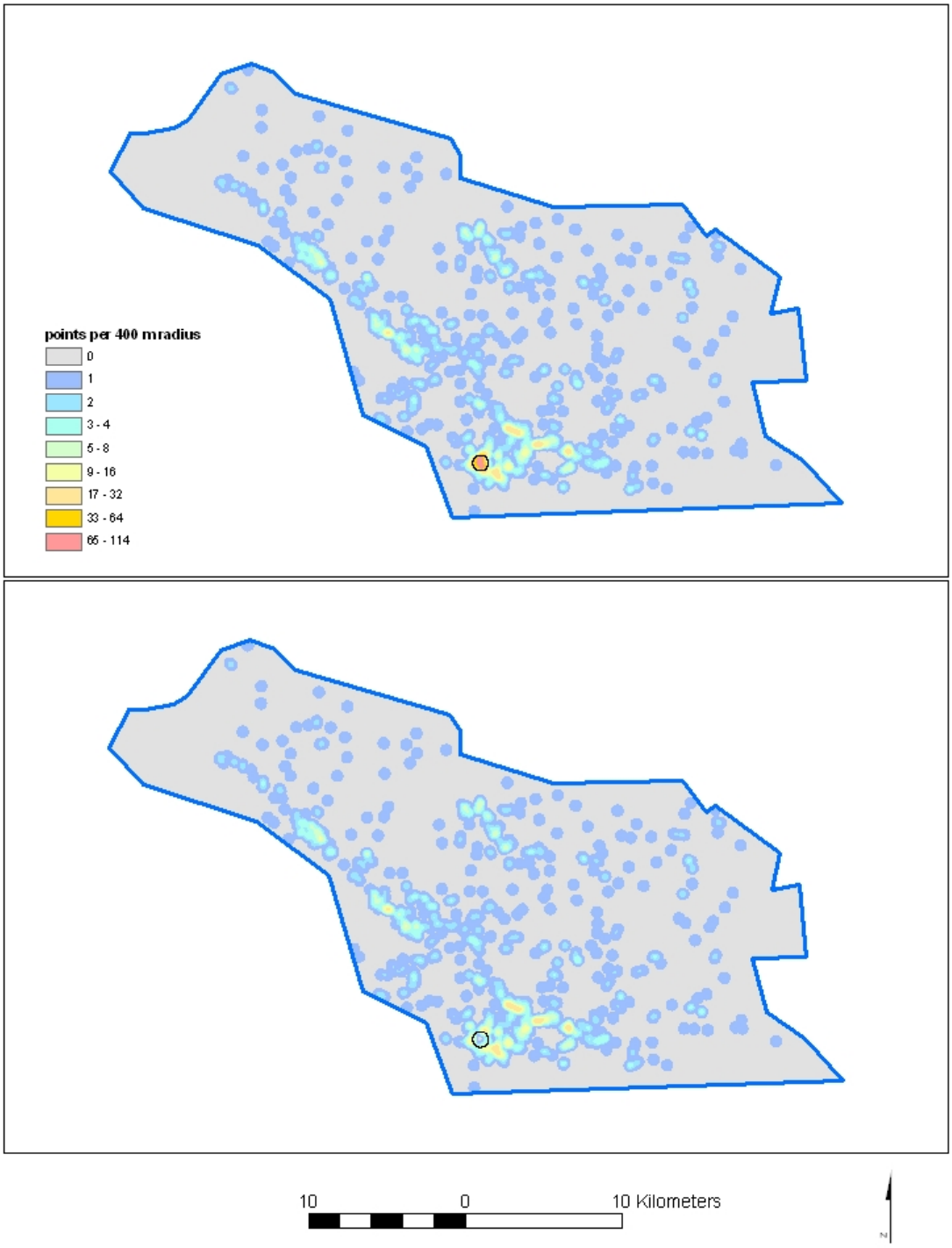


Figure 8. Density plots of the observed damage point dataset (top) and the declustered damage point dataset (bottom). Note the hot spot indicating a cluster in the southwestern corner of the study area (top; small circle on maps denotes the model cluster radius). This cluster is significantly reduced in the declustered dataset. Color scale is logrithmic.

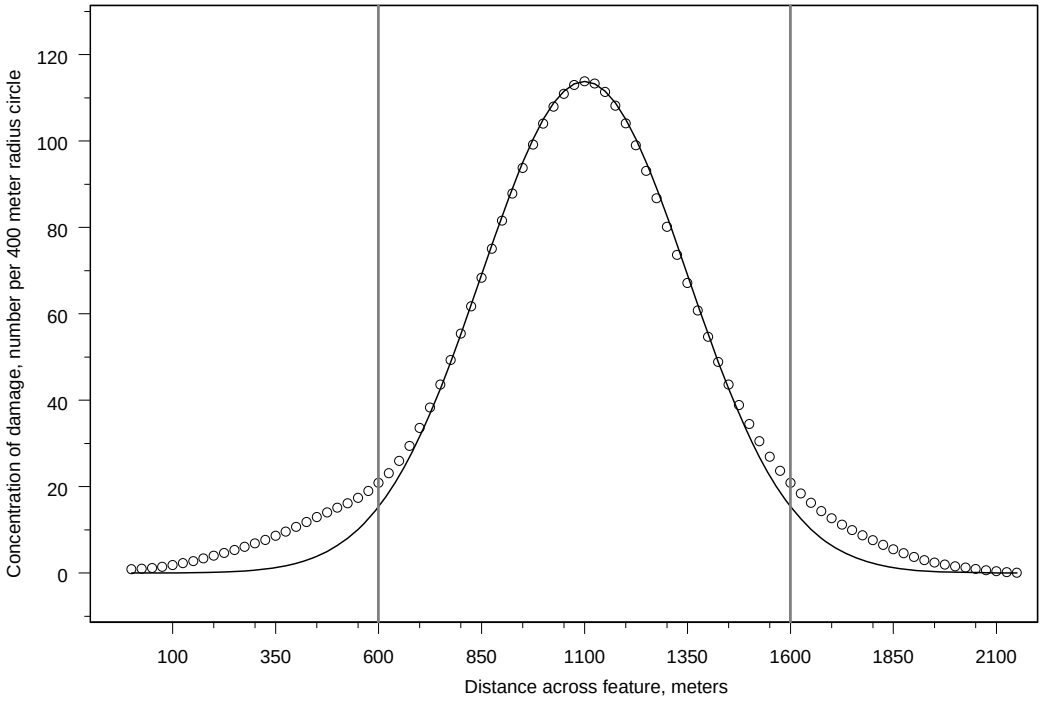


Figure 9. Circles are values from the density map taken along a cross-section. Black line is a gaussian curve (mean 1100, standard deviation 250) fit to the data. Note that the gaussian curve fits the data well to two standard deviations, indicating damage in this area forms a cluster, and little effect is seen beyond two standard deviations (grey vertical lines).

within the town of Los Gatos[2] and a radius of about 500 m. There are 310 points within this area, and they seem to be attributable exclusively to the cluster, with little effect seen outside the model's radius.

The 310 points within a 500 m radius of the center of the modeled cluster were removed, generating a dataset referred to for the remainder of this paper as the *declustered* dataset. The density plot of the declustered dataset is shown in figure 8 (bottom). The resulting dataset should be free of the effect of the dramatic clustering in the town of Los Gatos. However, it is possible that the clustering and the linear concentrations of damage are related to the same process as the lineaments, and that using the declustered dataset will reduce the sample size and introduce some bias. Since the nature of the cluster is not clear, both the original dataset and the declustered dataset will be examined in this report, and the results compared.

RESULTS: Association of damage with lineaments

As a first step, the point data is assumed to be homogenous over the lineament study area. To numerically approximate a random distribution of locations assuming a homogenous process (the probability is the same throughout the study area), the

[2] coordinates are (590285, 4120396) UTM zone 10 NAD27

lineament study area was rasterized at 5 m, and the distance of the center of each raster cell to the nearest lineament was recorded. This created a distribution against which to compare the observed distribution of damage.

In general the shape of these histograms is asymmetric and skewed to the right. This is caused by two competing effects: the increase of area with distance away from an object (the area between 1 and 2 radii from a point, for example, is less than the area between 2 and 3 radii from the point), and the confines of the study area. Thus these histograms will increase until the restrictions imposed by the study boundaries limit the area available at large distances from the object. For a circular study area and a single point at its center, the histogram of the area at a given distance from the point will increase until the radius is reached. For an irregular study area with multiple objects, the amount of area available within the study area (typically) gradually tapers off with increasing distance from the objects.

Figure 10 shows the histograms and density plots of the random and observed distributions. The median of the observed distribution, approximately 119 meters, is less than half the median of the random distribution, which is 240. Also note that the frequency of observed distances falls off much more rapidly towards the asymmetrical tail of the distribution than the random distribution. From this one can infer that the damage points are likely to be associated with the lineaments. In general, the damage points are simply closer to the lineaments.

A more formal statistical test can be performed to demonstrate that the distributions are different. Figure 11 shows the cumulative distributions for both the random and the observed data sets. Again, assume that the points within the damage dataset are independent of each other; and that there is no tendency of clustering or avoidance among the points themselves. Any spatial association between the damage and lineaments is interpreted to be due to external influences, and in this case tectonic processes are the likely candidate. A two-sided Smirnov test, which is based on the difference between the two cumulative distributions, confirms the two distributions are different at greater than the 99% confidence level. From the data it can be seen that the observed damage points are more closely associated with the lineaments than the random dataset, thus the two tend to be spatially related.

Now consider the inhomogeneous case, where damage is assumed to be homogeneous only within the area covered by streets. In the previous analysis it was assumed that the entire lineament study area was homogeneous, but this is not the case since the damage dataset only records damage in areas covered by streets. To take this into account the null hypothesis must be developed excluding the sub-areas that cannot contain damage to pavement and pipes breaks.

Using the approximation of the lineament study area provided by the scanned USGS map discussed in the previous section, the same calculations (finding the distance to the nearest lineament for each cell) were performed. The result is the distribution of the null hypothesis of spatial randomness under the condition that only areas covered by streets can contain a damage point. Figure 12 compares the two null hypotheses of randomness, that considering the entire study area and that considering only the area covered by streets. This comparison shows the difference between the two null hypotheses of complete spatial randomness, the first which considers the entire study area, and the second which considers an area restricted to the streets that occur within the

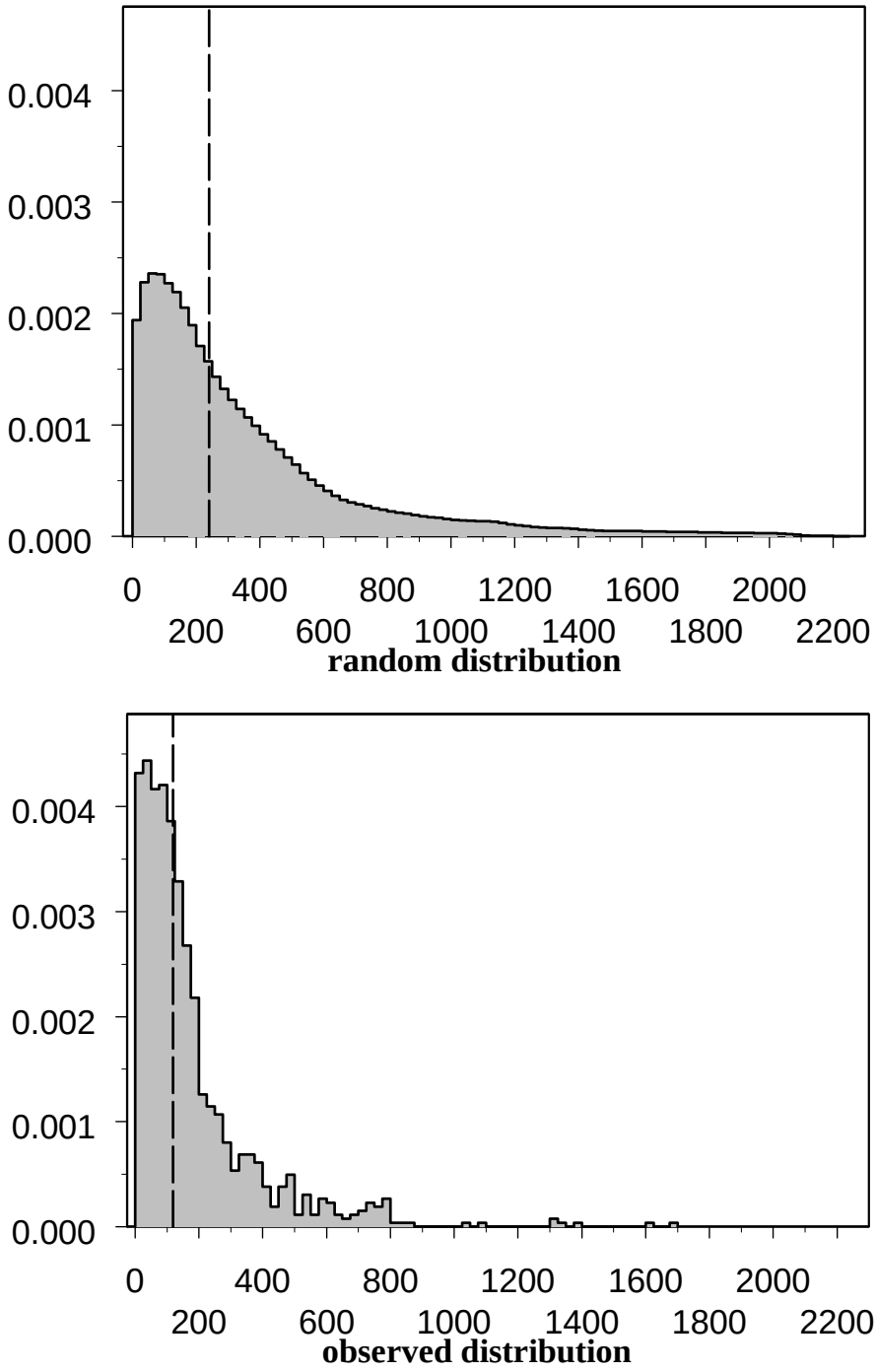


Figure 10. Histogram of the random and observed distributions of the distance to the nearest lineament, in meters. Dashed lines show the median.

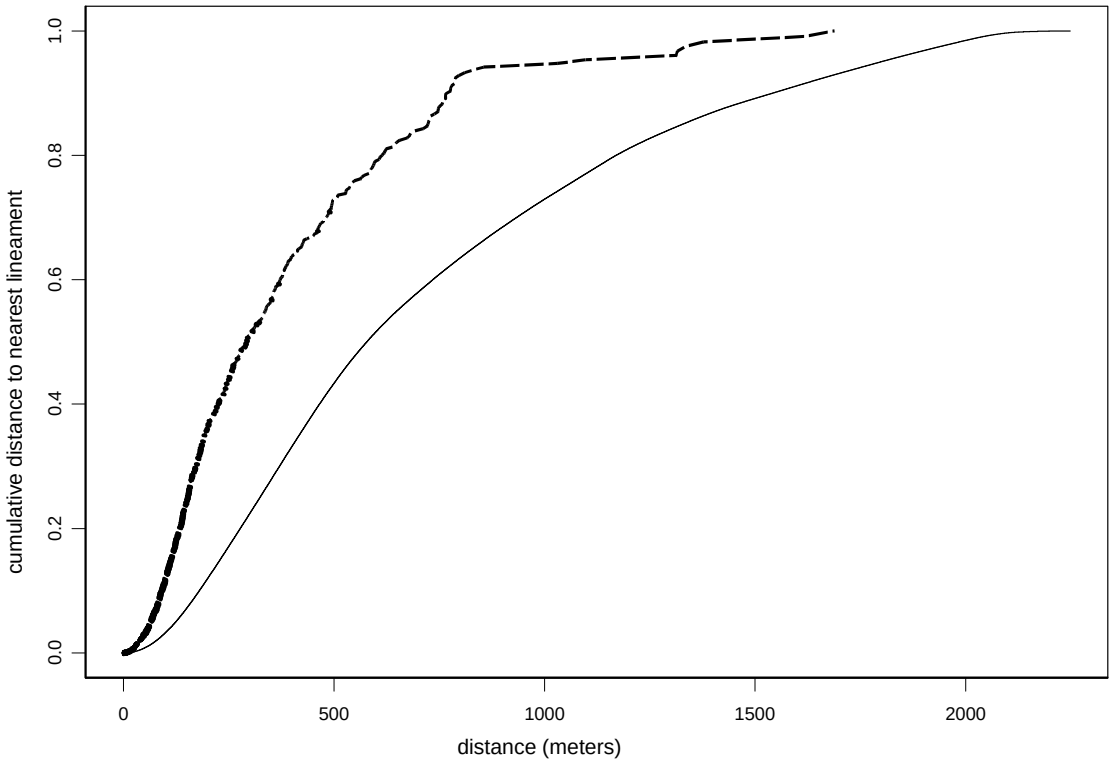


Figure 11. Cumulative distributions for the distance to the nearest lineament for a random set of points (solid, lower curve) and for the observed damage dataset (dashed, upper curve), for the lineament study area.

study area. The median of the latter distribution is 249, and the histograms are seen to match closely. Figure 13 shows a quantile-quantile plot comparing the distributions. In this case the difference between the two distributions is not great. Apparently the streets offer a robust sampling of the lineament study area, and the change to the distribution of the null hypothesis is minimal. Therefore the comparison of the observed damage to the null hypothesis of complete spatial randomness performed in the previous section is unchanged.

The previous analysis of comparing the observed damage to a null hypothesis of complete spatial randomness assumes that the cause of the damage is related to processes causing the lineaments, namely faulting along the range front of the Santa Cruz Mountains. If the large cluster of damage near Los Gatos, discussed in the previous section, is caused by a different process, say the focusing of seismic waves, then damage in this cluster could be adding bias to the statistical results. To address the question of whether or not the cluster is biasing, or even driving, the analysis, the analysis was repeated using the declustered damage dataset.

A repeat of the analysis confirms the original results, although the observed distribution for the declustered dataset is not quite as sharply peaked, as shown in the histograms in figure 14 where it is compared with the distribution for the homogeneous lineament study area. The median has shifted from 119 meters to 154 meters. This indicates that the clustering does have an effect. The median of the declustered data set is

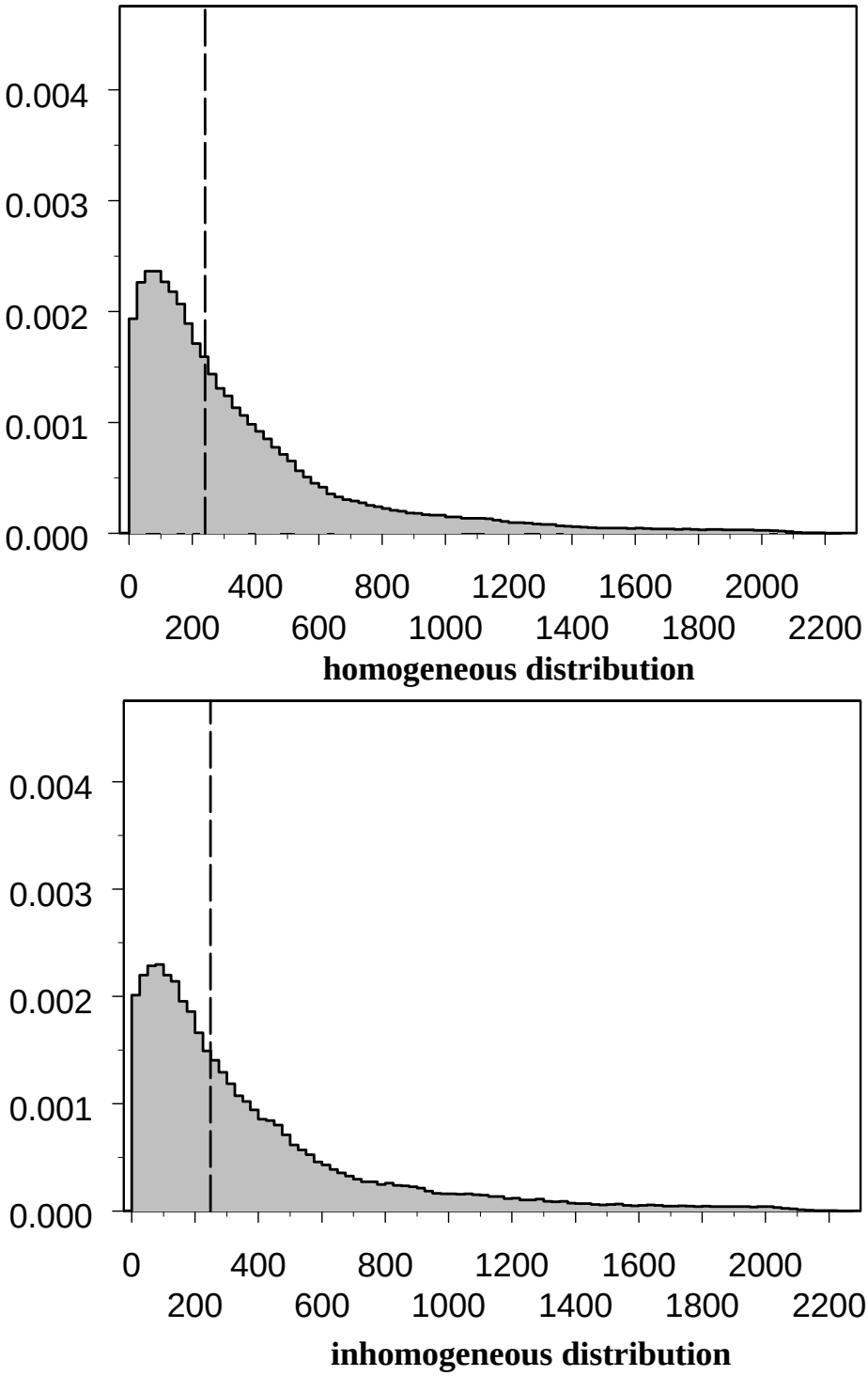


Figure 12. Histogram plots of the homogeneous (entire study area) and inhomogeneous (area covered only by streets) distributions of the distance to the nearest lineament, in meters. Dashed lines are medians.

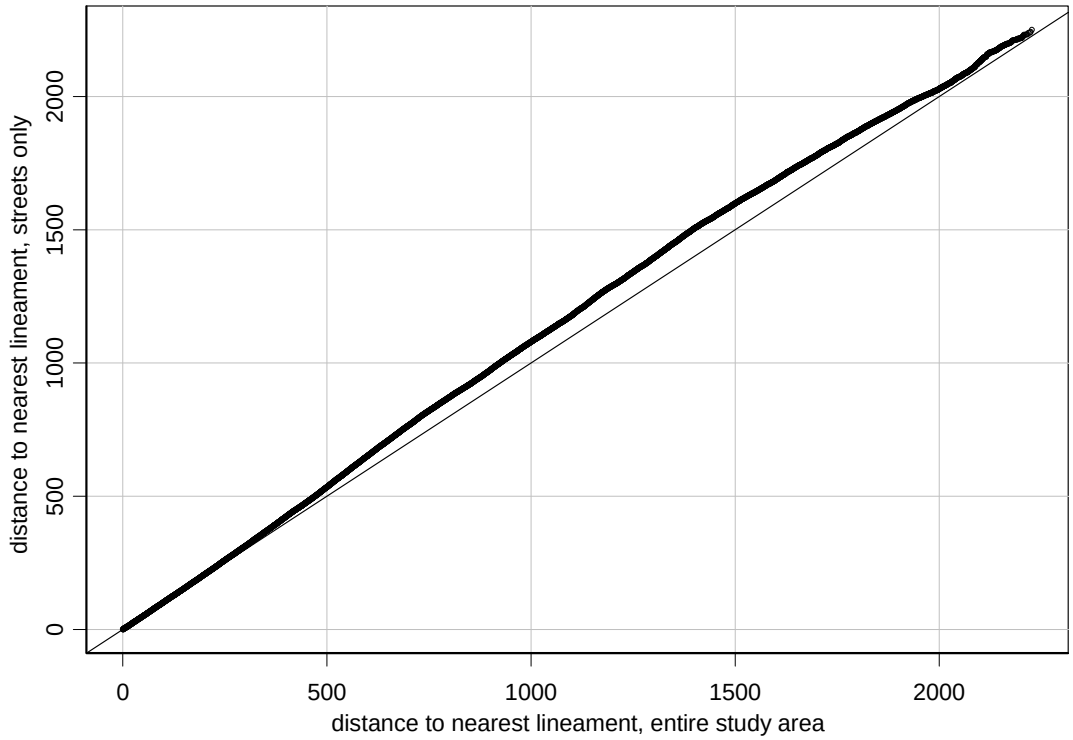


Figure 13. Quantile-quantile plot comparing the distribution of the homogeneous (entire study area) and inhomogeneous (streets only) distributions of the distance to the nearest lineament within the lineament study area. The distributions are very close to one another.

still closer to the lineaments than a randomly located set of points, but less dramatically so. The cumulative curves are shown in figure 15, and a two-sided Smirnov test again confirms that the distributions are significantly different from each other at the 99% confidence level. The damage points are located more closely to the lineaments than a randomly located set of points, even after the clustering effect has been mitigated and the inhomogeneity has been accounted for.

Investigating the lineaments by treating them as zones

Previous paragraphs have treated the lineaments as independent, individual lines. If instead one wishes to treat the lineaments as representing zones, then one must first convert the lineaments to zones, areas over which the process that generated the lineaments operates. Once zones are identified, points can be compared with them. Reasons for treating the lineaments or damage as zones might be that they are thought to be an expression of a stochastic process that operates over a zone, for which the existing lineaments or damage are one realization of that process.

To convert the lineaments to zones, one must first decide how to define the zones. A simple way would be to specify some distance, *d*, from the lineaments, less than which is defined as within the lineament zone. One could then buffer the lineaments at the distance *d* to create the polygonal zones. The damage could be examined to see if a majority was located within the lineament zones. However, this is a binary version (only two options are considered: less than distance *d*, and greater than distance *d*) of the

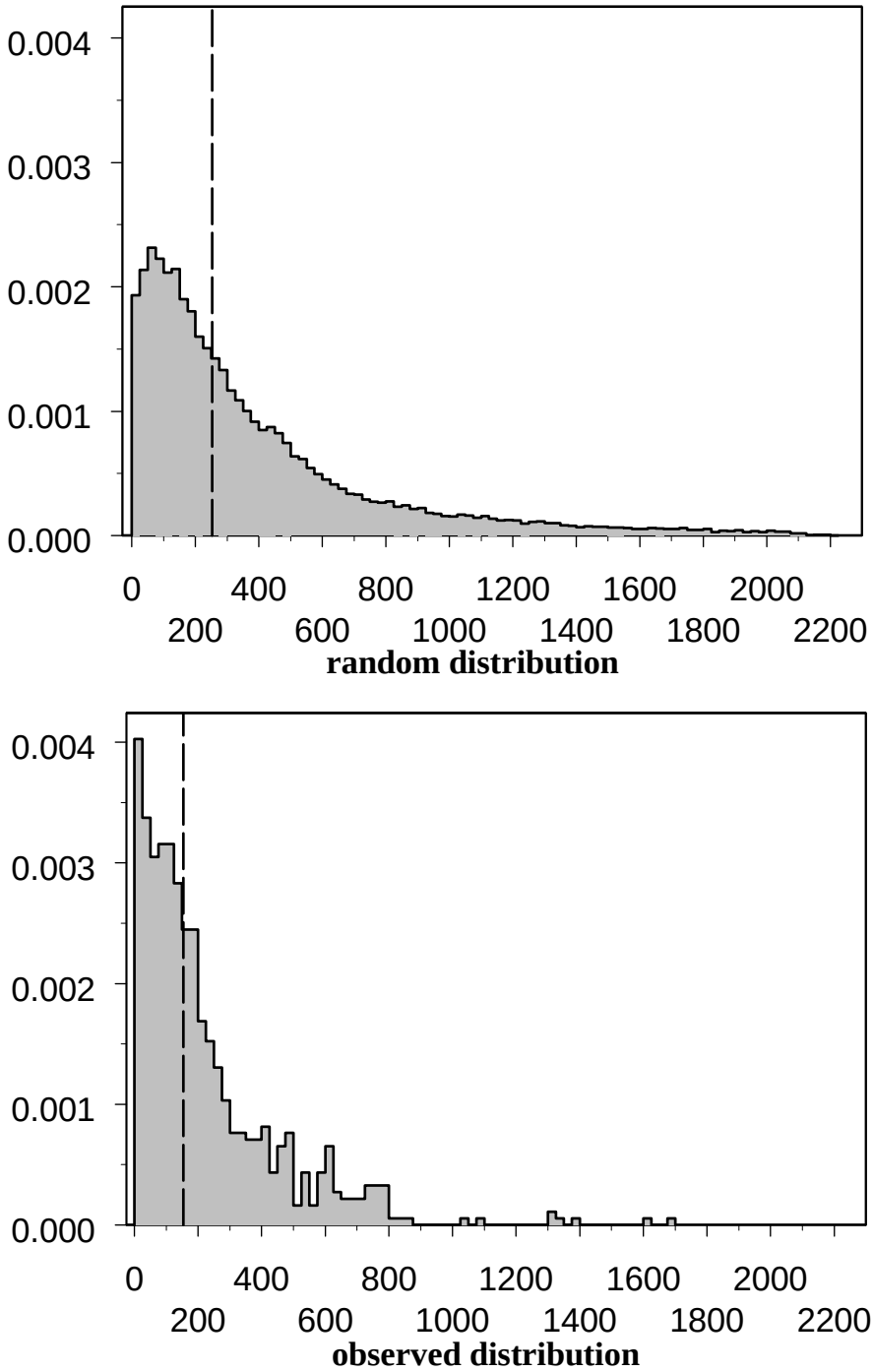


Figure 14. Histogram plots of the random and observed distributions of the distance to the nearest lineament, in meters, after removal of the damage cluster. Dashed lines show the median.

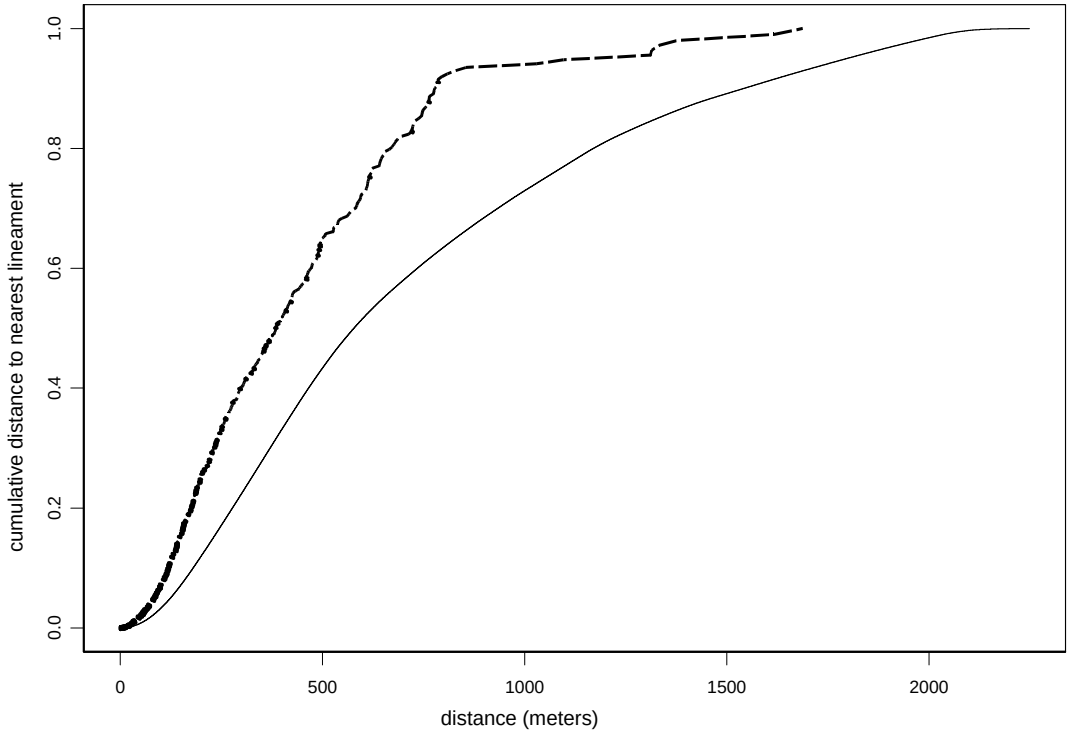


Figure 15. Cumulative distributions for the distance to the nearest lineament for a random set of points (solid, lower curve) and for the observed damage dataset (dashed, upper curve), for the declustered lineament study area.

original problem addressed in the previous section, where the entire range of distances of damage from the nearest lineament was considered. The binary case is a simplified version that considers only a single distance, *d*. Therefore if one defines the zones by means of a buffer, nothing is gained over the original analysis.

An alternative to buffering the lineaments is to calculate the lineament density, as was done for the damage density in the previous section. In the case of linear features, the total length of features within a specified window is calculated across the map. The units are length per area. Again, the analyst must determine the size of the window. The density of lineaments in the lineament study area was determined using a 400 m window and a cut-off density of one standard deviation above the mean lineament density. This was chosen to satisfy the author's judgment that the lineament zones should surround areas with abundant lineaments, yet not extend too far beyond the outer edge and into the adjacent empty space. Larger window sizes generate zones that extend beyond the concentration of lineaments into adjacent empty space, and smaller window sizes do not adequately combine dense areas into zones. Figure 16 shows the zones of lineaments defined as areas that are above the mean density of lineaments in the study area (top left), and as areas that are one standard deviation above the mean density of lineaments in the study area (top right). The first definition is compared with a buffer of 200 m from the lineaments (bottom left). This definition of a zone and the buffer are very similar, and are both judged to extend too far beyond the outer lineaments in the lineament groupings. The second definition of a zone, all areas one standard deviation above the mean lineament density, is deemed superior, and is shown with the lineaments (top right) and

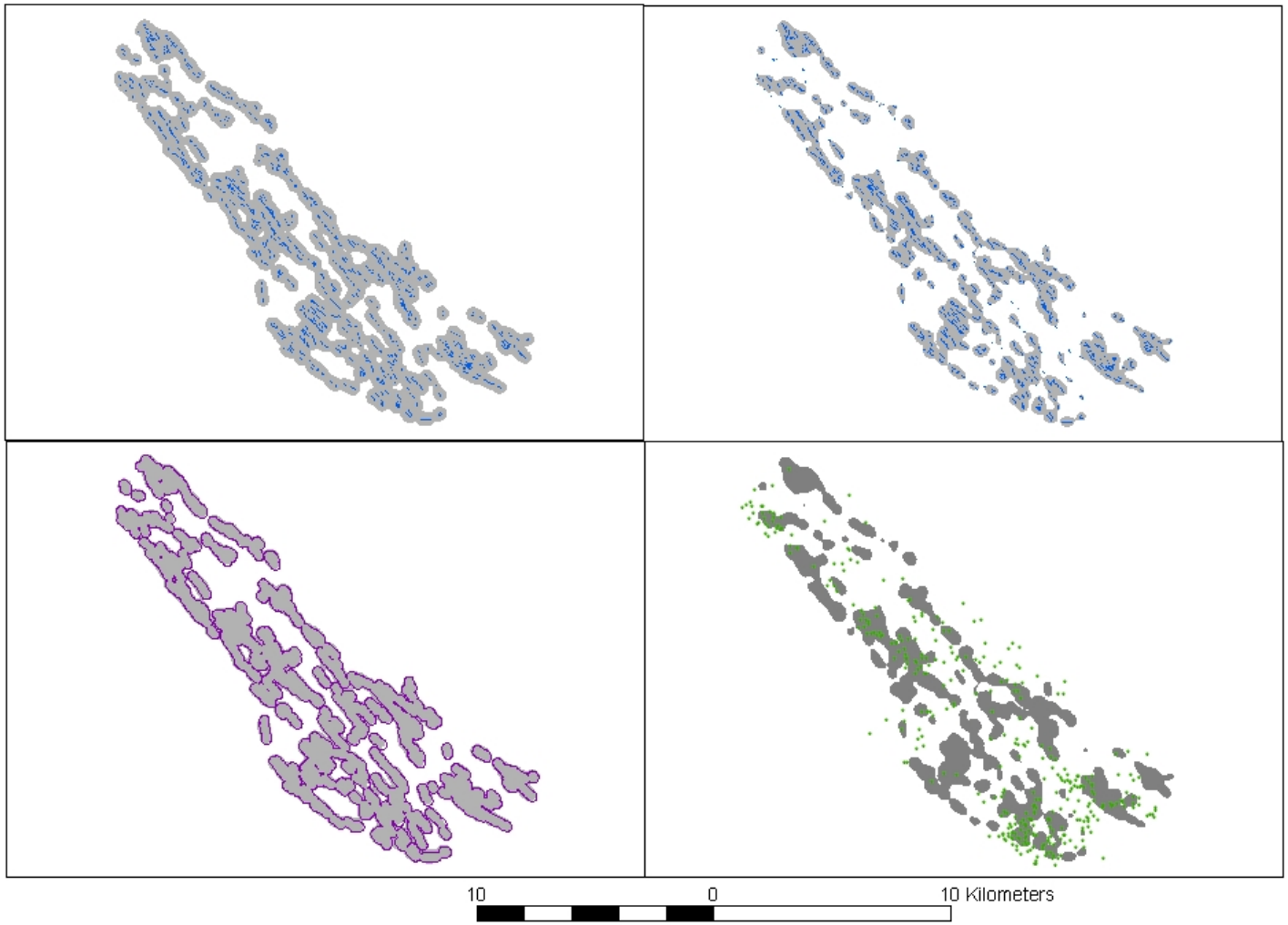


Figure 16. Alternate definitions of lineament zones (grey areas). Top and bottom left define the zones as greater than the mean density of lineaments (blue lines). This is very similar to a 200m buffer (purple) surrounding the lineaments (shown bottom left). Defining zones as greater than one standard deviation above the mean line density (right) preserves many interspaces between groups of lineaments and shrinks the edges of the zones, reducing the effect of the outer lineaments (top right shows lineaments (blue), bottom right shows damage (green circles)).

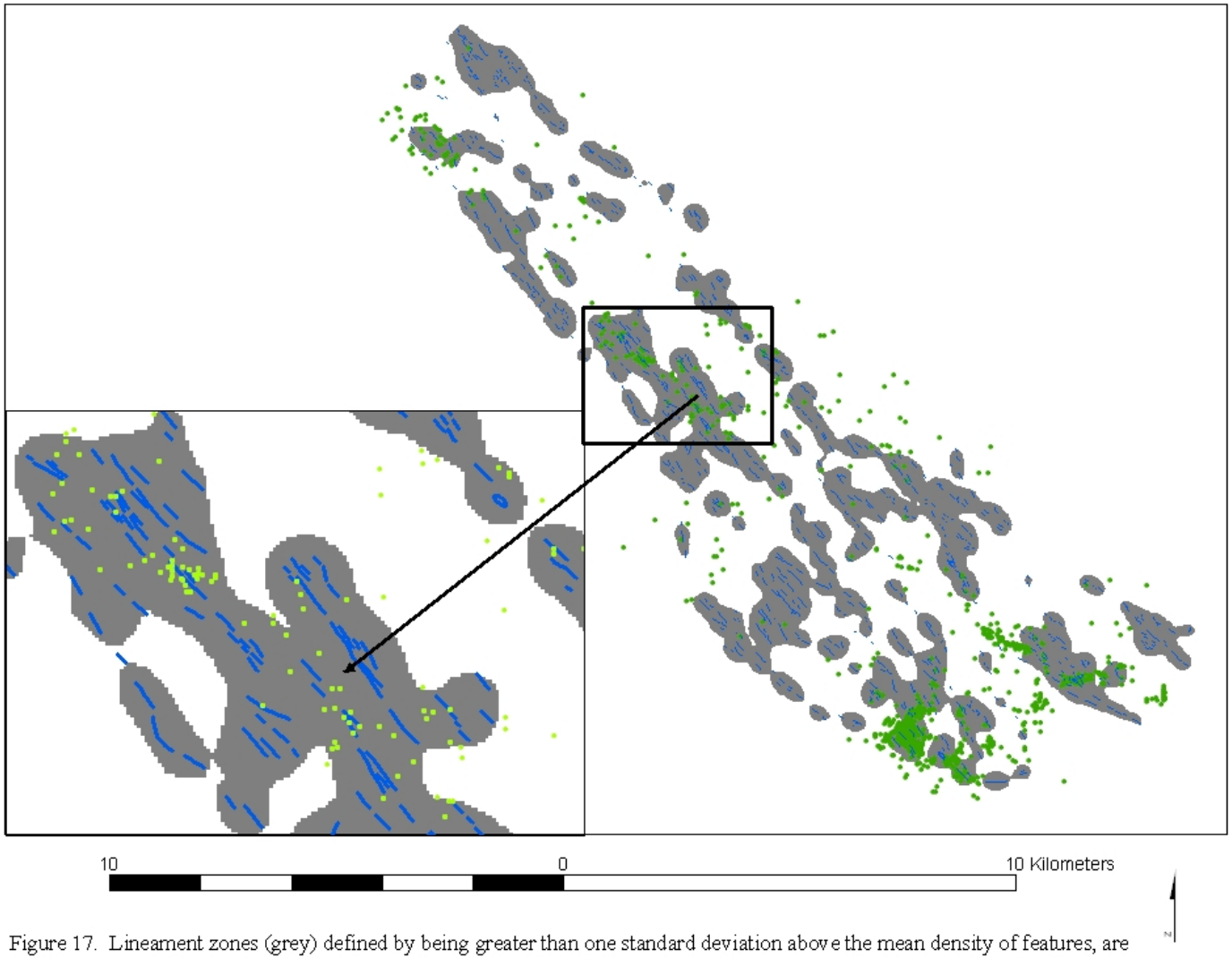


Figure 17. Lineament zones (grey) defined by being greater than one standard deviation above the mean density of features, are shown with lineaments (blue) and damage (green). Note how the damage falls primarily within the lineament zone in the detailed inset. The association of damage with the lineament zones, as defined, is statistically significant.

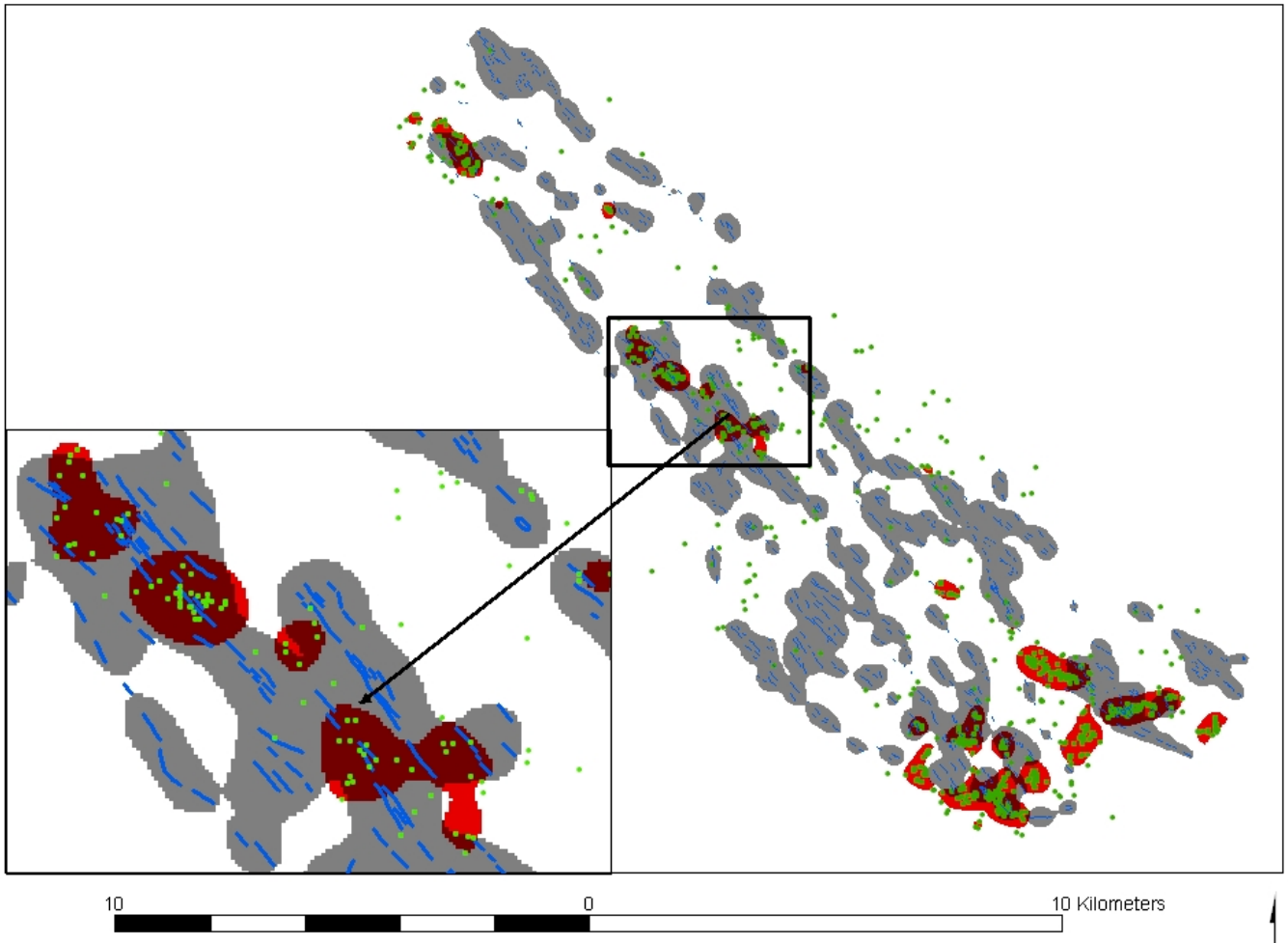


Figure 18. Lineament zones (grey) and damage zones (red) based on the declustered damage, defined by being greater than one standard deviation above the mean density of features. The zones largely overlap in the center, but only partially overlap to the northwest and southeast. Lineaments (blue) and damage (green) are also shown for comparison. Note how the damage falls within the lineament zone in the detailed inset.

the damage (bottom right) overlain. While the zones are ultimately defined based on the judgment of the analyst, the *procedure* used to define the zones is quantitative and repeatable. This allows for a quantitative comparison between datasets, something that is not possible if the procedure used to define the zone boundaries were based solely on judgment (as in a visual assessment).

The area where the density of lineaments is greater than one standard deviation above the mean within a 400 m moving window defines the zones of lineaments within the lineament study area. The area within the lineament zones makes up approximately 31% of the total lineament study area. The zones can now be compared with the damage data. Since, as determined in the previous section, the streets seem to offer a robust sample of the lineament study area, the complication of using an inhomogeneous study area is not considered in the following analyses.

One would expect that with 31% of the area covered, about 31% of the points should fall within the covered area. We can use the binomial distribution to test the hypothesis that the points are randomly distributed in space. Considering all damage that falls within the lineament study area, 652 out of 1046, or approximately 62%, of damage points fall within the lineament zones (figure 17). The probability that at least 652 out of 1046 points fall within the damage zone is essentially zero (less than 4.5×10^{-14}), thus we must reject the hypothesis that the damage is randomly distributed. The damage clearly shows a spatial association with at least some of the zones defined by the lineaments.

The same analysis can be performed using the declustered dataset, to mitigate the influence of clustering processes on the analysis, as discussed previously. For the declustered lineament study area, the lineament zones make up 30% of this area. In the declustered dataset there are 353 out of 735 damage points, or approximately 48% of the points, that fall within the lineament zones. This leads to 222 expected points within the lineament zones under the assumption of complete spatial randomness (based on the binomial distribution). Again, the probability of 353 or more damage points falling within the damage zone is essentially zero (less than 6.4×10^{-14}), and the damage clearly shows a spatial association with at least some of the zones defined by the lineaments.

One can also compare the lineament zones with zones of damage generated from the declustered dataset, where the zones for the damage are defined using the same parameters: a 400 m search radius, and a cut-off of one standard-deviation above the mean. The zones are overlain in figure 18. The lineament zones make up approximately 30% of the lineament study area, while approximately 53% of the area of the damage zones overlies the lineament zones. While a formal statistical analysis of the probability of the proportion of irregular shapes overlying one another within an irregular area is beyond the scope of this paper, a few qualitative observations can be made regarding the overlapping zones. The first is that 53% overlap seems rather significant, because under complete spatial randomness for a set of points one would expect about 30% overlap. The second observation is that the area of overlap shown in the inset seems to be the most significant overlap in the lineament study area. No clear linear pattern exists across the study area to suggest damage follows along a particular set of lineaments that could be inferred to be a fault. Rather, the zones overlap in patches, with perhaps the suggestion of a weak linear trend from southeast to northwest.

Key decisions in the flow of the analysis in this section are based on the judgment of the analyst. These judgments include optimal window sizes and cut-off values used in

generating the density plots, and the method used to decluster the data. In this sense the analyses are exploratory in nature. However, as emphasized, the quantitative *procedures* defined for the analyses provide a quantitative foundation for repeating, and therefore comparing, analyses among different datasets. Furthermore, seeking a quantitative procedure by which to define fuzzy concepts such as "lineament zones" or "clusters of points" helps focus attention on the logic behind the scientific intuition that is often used to guide analysis. It is hoped that the procedures discussed have this effect.

METHOD: comparing the damage data to complete spatial randomness, and looking for alignment of damage

The previous section has suggested that the damage is preferentially located in zones similar to those defined by mapped areal photographic lineaments. The following sections are meant to present general methods of analyzing a point pattern, of which the damage data is one example. The question of randomness of the damage data will be explored without considering the previous results so that these new methods may be illustrated.

The map pattern of the damage in the Santa Clara Valley from the Loma Prieta earthquake appears to be concentrated in a linear band outboard of, and parallel to, the range front of the Santa Cruz Mountains. There are also quite a number of damage points scattered throughout the damage study area. Is there a quantitative test that can be brought to bear that rules out the criticism that the damage points are randomly located? Can linear structure be quantitatively defined within the scattered damage points?

To answer these questions a model of randomness (null hypothesis) must be constructed against which to compare the observed data. For a sample of points distributed over an area, what is meant by an alignment of points? One interpretation is that line segments defined by pairs of points tend to be aligned preferentially. Points clustering about an imaginary line would result in pairs of points aligned in a particular direction over a range of distances, whereas points scattered about an area would result in pairs of points with no particular alignment, given the constraints of the study area. Figure 19 shows two synthetic datasets defined over a 1 km by 1 km region. The first point set consists of ten points randomly located within the area, defined by combining ten random numbers, from 0 to 1, for the x-coordinates with ten random numbers, from 0 to 1, for the y-coordinates. The second point set consists of ten aligned points, defined by combining ten random values, from 0 to 1, for the x-coordinates and assigning the y-coordinates according to the equation $y = 0.6x+0.2$ (an arbitrary line within the study area). Each pairing of points within the study area defines a vector, and the distribution of the directions defined by vectors, referred to from here on as *direction vectors*, can be used to examine structure within a point set; namely, whether or not the points are randomly distributed throughout the study area, and whether or not other structure can be detected. *Direction vectors* are simply the vectors defined by the pairs of points in a point set, and have length and direction defined as for any vector. *Direction*, in this case, refers to the compass direction in which the vectors point; the head and tail of the vectors are less significant for this analysis. In this paper the focus is on the orientation component of the direction vector and not the length.

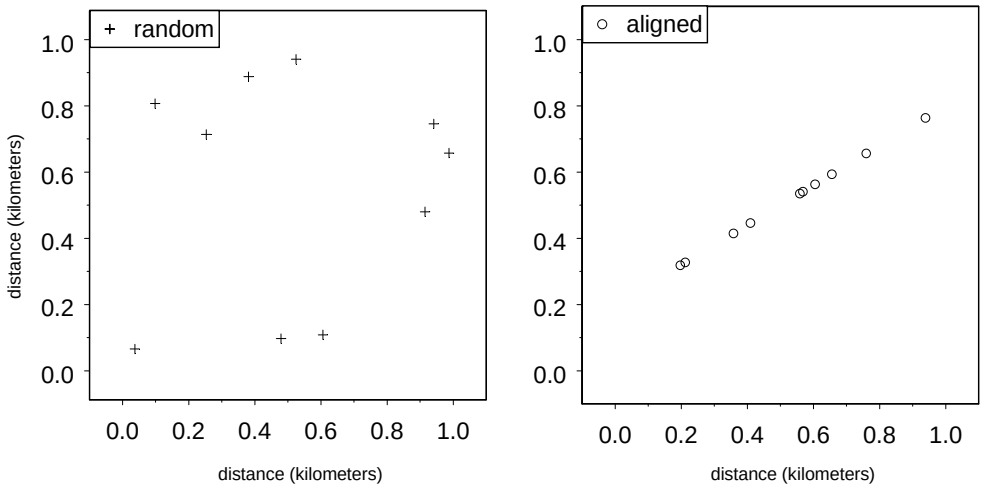


Figure 19. Location of 10 randomly placed points (left) and ten points placed randomly along a line (right) according to y=0.6x+0.2 (an arbitrary line). The direction vectors connecting the random points will not show a preferential direction; the direction vectors connecting the aligned points will.

Figure 20 shows the distribution of the direction vectors for both the random point set and the aligned point set shown in figure 19. The distribution has been quantized into 5º bins and the bin mid-points connected by lines instead of using histogram bars. This emphasizes the shape of the distributions, and allows them to be added together easily. Note how the random point set results in direction vectors distributed roughly uniformly, whereas the aligned point set produces direction vectors that are aligned in one direction. Strictly speaking, the study area must be circular for the random distribution of the direction vectors to be perfectly uniform, but practically the deviation from a uniform distribution caused by a square study area is less than 5 percent. A Monte-Carlo test of 100 simulations and 2000 points revealed that distributions of both the circular and square study area were well within each others' respective q05 and q95 (quantile) interval.

For a given point set with *n* points, and defining the direction vectors as any pairing of points other than a point with itself, there are $n(n-1)$ total direction vectors. Half of these direction vectors are equivalent but of opposite sign e.g. direction vectors for points P_1 and P_2 are defined as $P_1 - P_2$ and $P_2 - P_1$. For some analyses it may be useful to consider the full set of direction vectors. Here north is defined as 0 degrees, and only the direction vectors between -90 and 90 (northerly) are considered, of which there are $n(n-1)/2$ in the point set.

A set of randomly scattered points will, more or less, fill a study area. The shape of the study area, however, may limit the amount of scattering possible for a set of randomly located points. The direction vectors will therefore be biased by the shape of

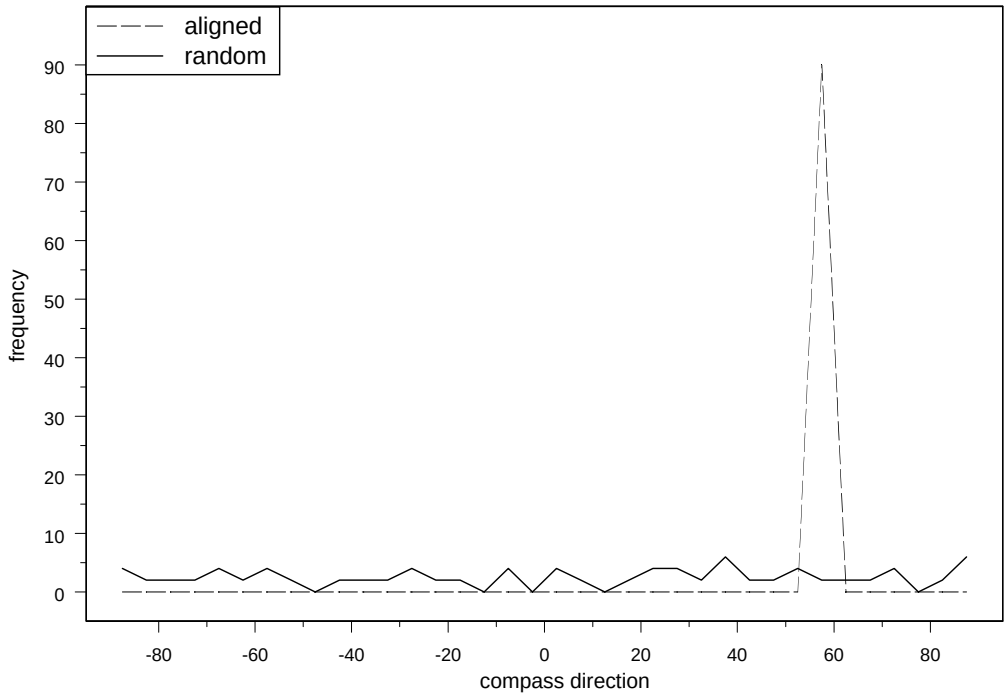


Figure 20. Frequency of direction vectors, in 5 degree intervals of compass direction, for the random point set and the aligned point set. The direction vectors of the random point set show a roughly uniform distribution, whereas the direction vectors of the aligned point set show a peaked distribution in the direction of the line along which the points fall.

the study area. A circular area containing randomly scattered points produces no preferential direction for pairs of points, on average, and the resulting distribution for pairs of points is the uniform distribution. However, for an irregular area the random distribution of direction vectors is not so straightforward. This is because any irregularly shaped area may afford the opportunity for more points in one sector, biasing the resulting direction vector distribution. For irregularly shaped areas methods must be developed for defining the random distribution of direction vectors.

One way to accomplish this is by Monte Carlo simulation: generate random points within the study area, measure the distance and direction for all pairs of points, repeat the experiment many times and average the result. This approach involves some programming and a fair bit of computer time, but has the advantage of producing an experimentally robust probability distribution where confidence intervals can be estimated.

Another approach is to calculate the probability using a numerical approximation of the theoretical solution. This approach is advantageous because it is straightforward and rapid once the method is understood. This approach is best explained in two parts. The first part involves finding the probability density function for a fixed point paired

with each member of a randomly located point set within a study area, and the second involves combining many of these into an overall probability density function for the entire area.

Consider a fixed point, P_1, within the study area (figure 21). Within the study boundary there is a fixed amount of territory that is between angle θ and angle $\theta + h$ from P_1. This territory consists of two wedge-shaped areas opposite each other on either side of P_1. A point, P_i, added to either of these wedge-shaped areas would result in a direction vector, defined by the point pair (P_1, P_j), having an angle of between θ and $\theta + h$, $\theta < \angle(P_1, P_j) < \theta + h$. If the second point, P_j, is added to the study area at random, it is as likely to land at one location within the study area as any other. Therefore the probability of a point landing in the area defined by the wedges is equal to the area of the wedges divided by the total area under study. This is shown in figure 21 as $\frac{a_1}{A}$, where a_1 is the combined area of both wedges and *A* is the total study area. Dividing the study area up into *n* wedges for *n* desired slices of *h* degrees each and recording the ratio of $\frac{a_j}{A}$ for each point pair (P_1, P_j) will result in a probability density function for the fixed point,

P_1.

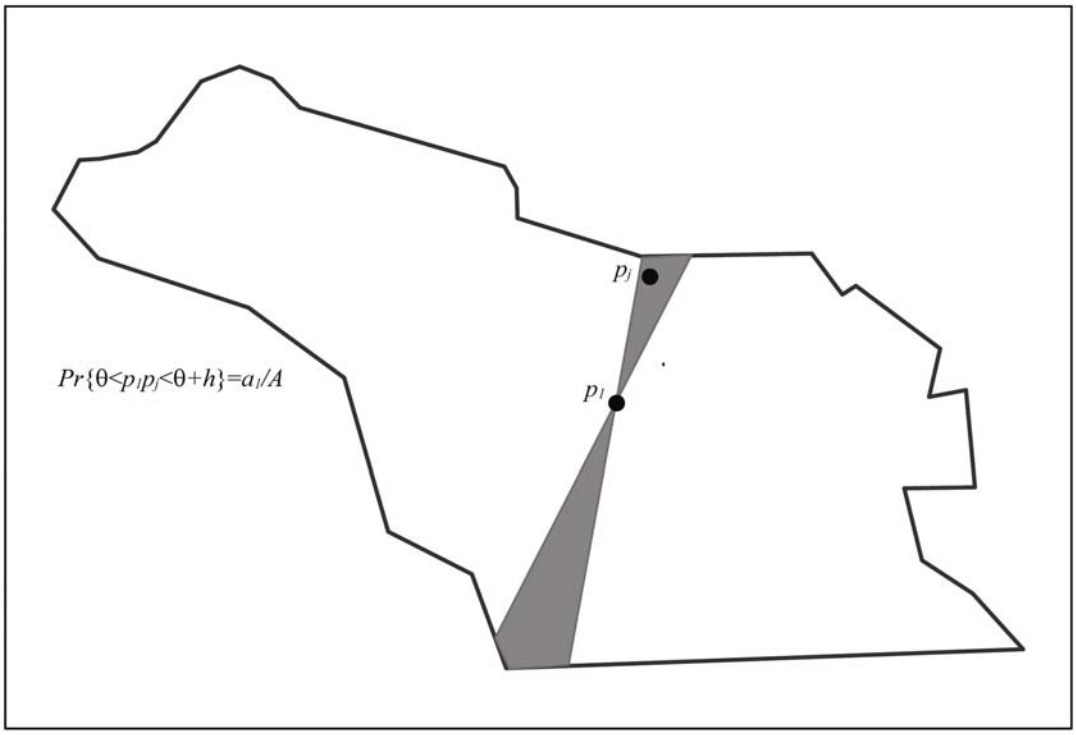


Figure 21. Given a fixed point, p_1, in the damage study area (outlined in black), the probability that a randomly located point, p_j, will form a direction vector, p_1p_j, with a compass direction of between θ and θ+h is the area of the sub-region that is within the stated compass direction (dark grey) divided by the total area of the study area, $Pr\{\theta<p_1p_j<\theta+h\}=a_l/A$.

This probability density function can be numerically approximated by estimating the area of each wedge. This is accomplished by rasterizing (representing the study area by a finite number of pixels) the study area and counting the pixels (which have a known area) between θ and $\theta + h$ degrees from P_1, for all cells in the raster dataset. So, by rasterizing the study area, a probability density function for the angle from a fixed point, P_1, can quickly be constructed by comparing P_1 to every other cell in the raster dataset.

The second part of the explanation extends this idea to all points in the study area. Let each raster cell (pixel) within the study area represent a point, denoted P_i or P_j. For each point, P_i, consider the direction vectors for all paired points $(P_i,P_j), \ni P_i \neq P_j$, in the study area. This yields the probability density function for the angle of a direction vector created by adding a new point, P_i. Summing across all points P_j within the study area yields the probability density function for the angle of direction vectors created by all pairs of a set of random points (P_j, P_i) within the study area.

An approximation of the probability density function for the direction vectors of pairs of random points within the irregular study area can be constructed by rasterizing

the study area at the desired level of precision and calculating the directions for all pairs of points. Since each raster cell has the same area, the units cancel and one need only consider the number of points within a given direction to the total number of points within the study area.

Now that a method exists to generate a probability density function of the direction vectors for a randomly located set of points within an arbitrarily-shaped area, the resulting theoretical distribution can be compared to an observed distribution. Since both distributions are circular and non-gaussian, a Chi-square test based on circular ranks test can be employed to see if the distributions differ significantly (Fisher, 1995).

RESULTS: comparison to complete spatial randomness and alignment of damage

To test whether the damage points are non-randomly located, a distribution of the direction vectors for randomly scattered points was compared to the empirical distribution of direction vectors for the study area. This test is comparing the damage data with complete spatial randomness, so the entire dataset, rather than the declustered dataset, should be used. The reason for this is that the relation of damage to one or more processes would legitimately indicate a departure from random behavior; additional processes controlling damage locations does not affect the outcome of the test. The distribution of direction vectors for a random set of points was constructed using the numerical approximation method discussed in the previous section. A histogram of the direction vectors, with a density line superimposed, is shown in figure 22 (bottom). Note how the shape of the study area causes a preferential alignment in the northwest direction.

The distribution of direction vectors for the observed damage locations are shown in figure 22 (top) as a histogram and density plot. There are over one million direction vectors for the sample of 1427 damage locations. Note that the two histograms differ somewhat in shape, but more importantly the mode in the random distribution is displaced from the mode in the observed distribution.

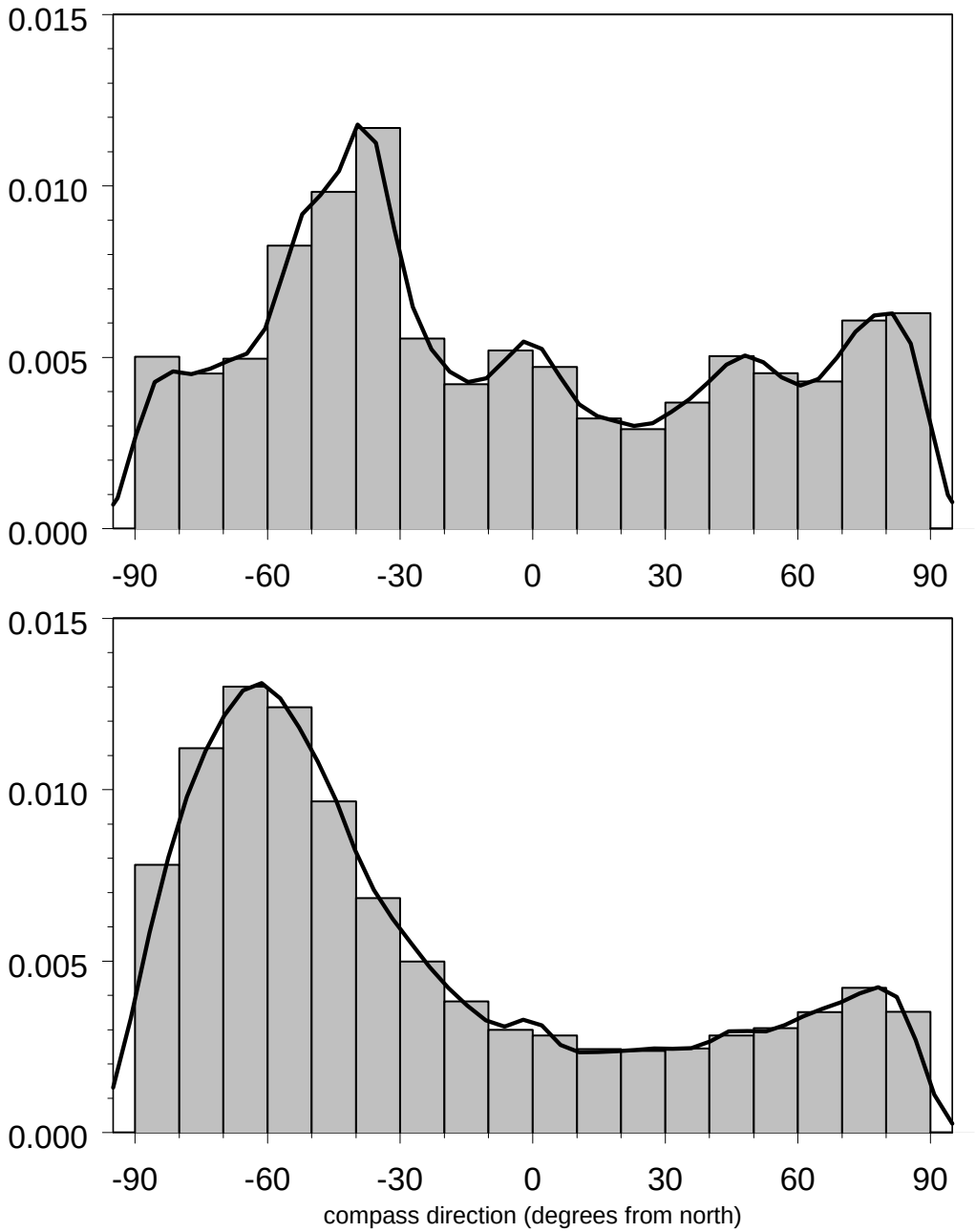


Figure 22. Combined histogram and density plots (black lines) for the observed distribution of direction vectors (top) and the theoretical distribution of direction vectors (bottom). Note that the observed distribution is multimodal, whereas the theoretical distribution is (largely) unimodal. The y-axis is based on the density plots, which normalize the area under the density curve to one.

Figures 23 and 24 compare the distributions to each other, using a quantile-quantile plot and cumulative plot, respectively. As the quantile-quantile plot (figure 23) shows, the two distributions differ noticeably, with the observed distribution having much fewer direction vectors in the range of -90º to about -50º than the random distribution. This difference also stands out on the cumulative distribution plot (figure 24), which is the distribution that is used for the Chi-square test based on the circular ranks test to compare two distributions. Judging visually, there is a clear distinction between the two distributions.

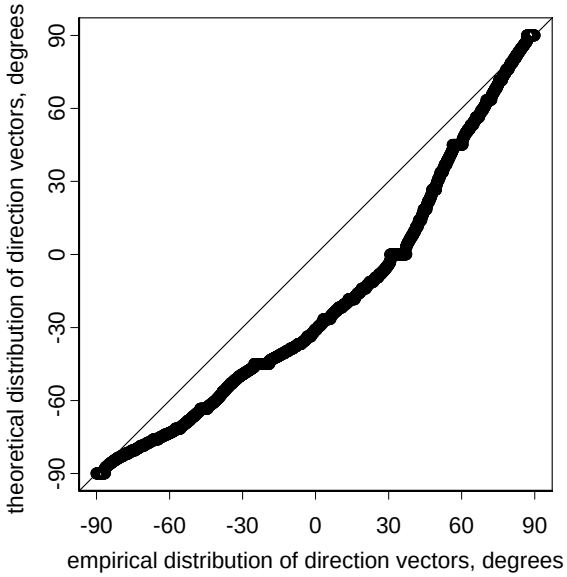


Figure 23. Quantile-quantile plot comparing the observed and theoretical random distributions formed by the direction vectors.

The non-parametric Chi-square test based on circular ranks was performed to quantify the difference between the two distributions (Fisher, 1995). The test showed a highly significant difference, beyond the 99% confidence level. However, such tests can be sensitive to very large sample sizes. For example, Rock (1988) points out that the significance levels for large (>100) samples are approximated for the Smirnov test, the non-parametric test used to distinguish two different distributions in the previous section, so supporting evidence for the degree of difference between the distributions would be helpful.

A Monte Carlo simulation that placed points within the study area at random locations was performed to test the veracity of the numerical approximation method and to confirm the results indicated by the Chi-square circular ranks test. 1427 points, equivalent to the sample size of the damage observations, were distributed at random locations throughout the study area, and the distribution of the direction vectors calculated. 100 simulations were performed, and the average, 0.05, and 0.95 quantiles recorded at five degree intervals. Figure 24 displays the results of the Monte Carlo

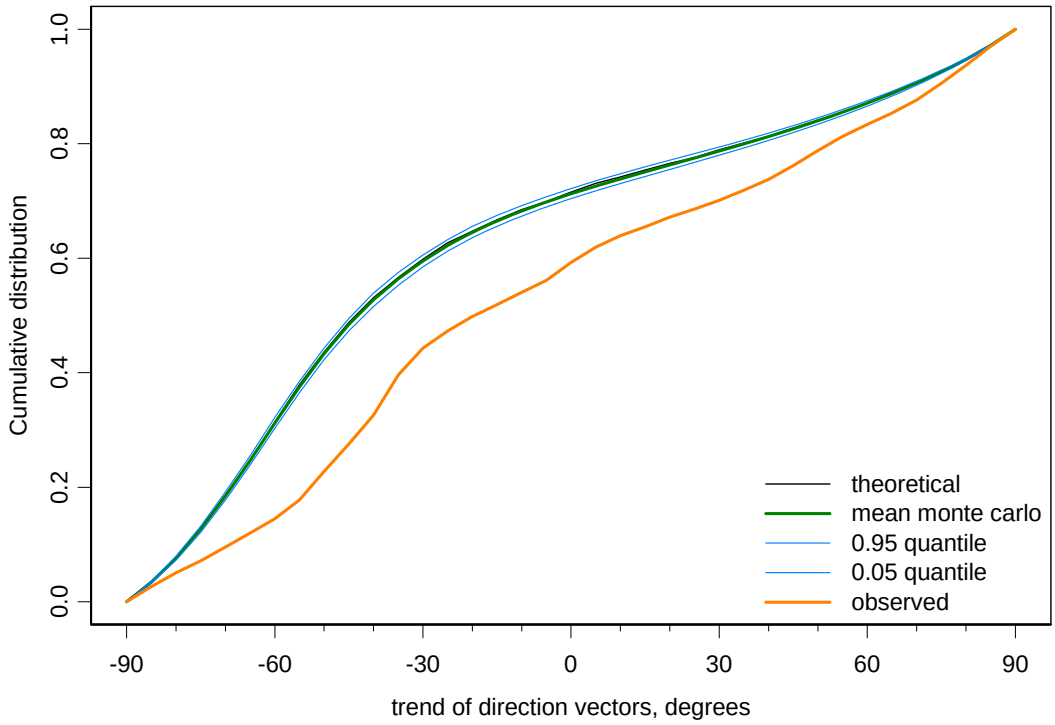


Figure 24. Cumulative distributions of the direction vectors for the theoretical random, mean monte carlo random (including the 0.05 and 0.95 quantile confidence intervals), and the observed damage data set. The mean monte carlo and theoretical random distributions are almost indistinguishable in this plot.

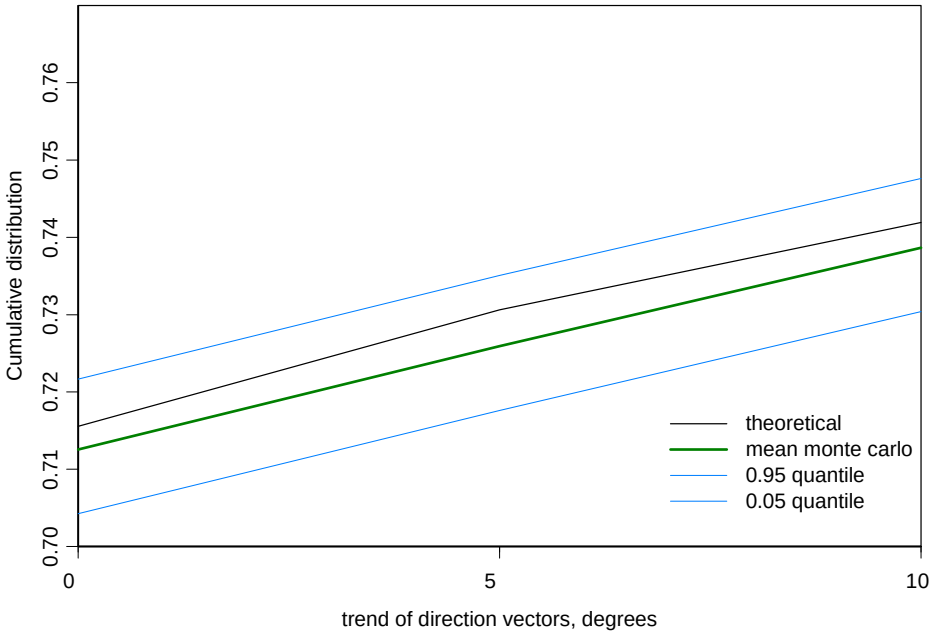


Figure 25 (close-up). Close-up of figure 25, showing that the difference between the theoretical distribution calculated numerically is close to the mean monte-carlo simulation, and within the 0.05 and 0.95 confidence envelopes.

simulation, including the 0.05 and 0.95 quantiles that serve as confidence envelopes. The distribution of the direction vectors for the average Monte Carlo simulation and the numerically calculated distribution are indistinguishable in the figure, and upon close inspection show only very minor deviations from each other (figure 25 shows a close-up of the graph, to better distinguish the numerically-calculated theoretical distribution of complete spatial randomness and the mean Monte Carlo distribution of complete spatial randomness). This indicates that the numerical solution is a viable alternative to Monte Carlo simulation for this type of analysis. The 0.05 and 0.95 quantile confidence envelopes are also quite close to the average distribution, indicating that there is little variation within the random distribution of direction vectors, at least with the sample size given. The observed distribution is clearly outside of the 0.05 and 0.95 confidence envelopes, supporting the results of the Chi-square circular ranks test. The damage data is clearly not randomly distributed.

Three potential problems with the investigation must be mentioned. The first is preferential sampling due to the limitation of damage being confined to public areas, primarily streets. Since streets are commonly arranged in grids, there is the danger of sampling finely parallel to the streets and coarsely at an angle to them, introducing anisotropy into the sampling process. However, a street map of the study region revealed no such systematic pattern to the street network, and the previous investigation showed minimal bias in the lineament study area. The study area is large enough to encompass several communities, each with separate street patterns that do not coincide in direction with each other. A related problem is that sampling along streets can introduce a linear pattern in the data for small local areas. This is noticeable within the large cluster describe in the previous section. The third potential problem is the possible association of damage with surface geology. While there is undoubtedly a relationship between the

two, the surficial geology is also closely related to the tectonics of the area, with alluvial fan units forming along the mountain front and giving way to broad mudflats towards the shore of the bay. This creates a trend in the spatial arrangement of the Quaternary geologic units. The damage points do not appear to be preferentially associated with any particular geologic unit. Indeed, they seem to cut across the grain of the Quaternary geology. Since there is no obvious relationship between the two independent of the local tectonic influence, and the two likely co-vary because of this, it is assumed that the Quaternary geology has an insignificant effect on the location of the damage.

The results indicate that the observed distribution of damage in the Santa Clara Valley differs from that of randomly distributed locations. Furthermore, the observed distribution shows distinct peak at the mode, 40º northwest (figure 22). This peak is narrower than the broad mode of the random distribution. This suggests preferential alignment of points at 40º northwest.

The numerical method used to test the null hypothesis of a random distribution of direction vector orientations is a viable alternative to Monte Carlo simulation, and may be preferred because it is easy to calculate.

DISCUSSION: Development of a process model and a method of spatial filtering

Visually the pattern of damage in Santa Clara Valley seems to be preferentially aligned in a northwest direction along a single, major line near the range front, with perhaps some smaller, secondary alignments further from the range front, and some scattered points throughout the study area (figure 2). The density plot (figure 8, top) also shows that a significant cluster appears in the dataset near the town of Los Gatos and the distribution of direction vector orientations for the damage locations also shows a prominent spike at approximately 40º northwest. It appears at least three patterns are combined within the damage data. The damage forming the cluster pattern was separated previously. Is there a quantitative model that can be applied to the dataset to separate the declustered damage into two subsets, one representing a linear pattern and the other a random pattern?

Visual inspection of the declustered damage, and the distribution of the resulting direction vector orientations, suggest a simple end-member model for damage, namely that damage is either aligned in a northwest direction, or scattered randomly throughout the study area. This combination of patterns would result in the direction vectors being defined by a mixture of pairs of random points, pairs of aligned points, and pairs of one random and one aligned point.

To investigate this behavior, consider the synthetic dataset of random and aligned points from figure 19, combined into one plot in figure 26.

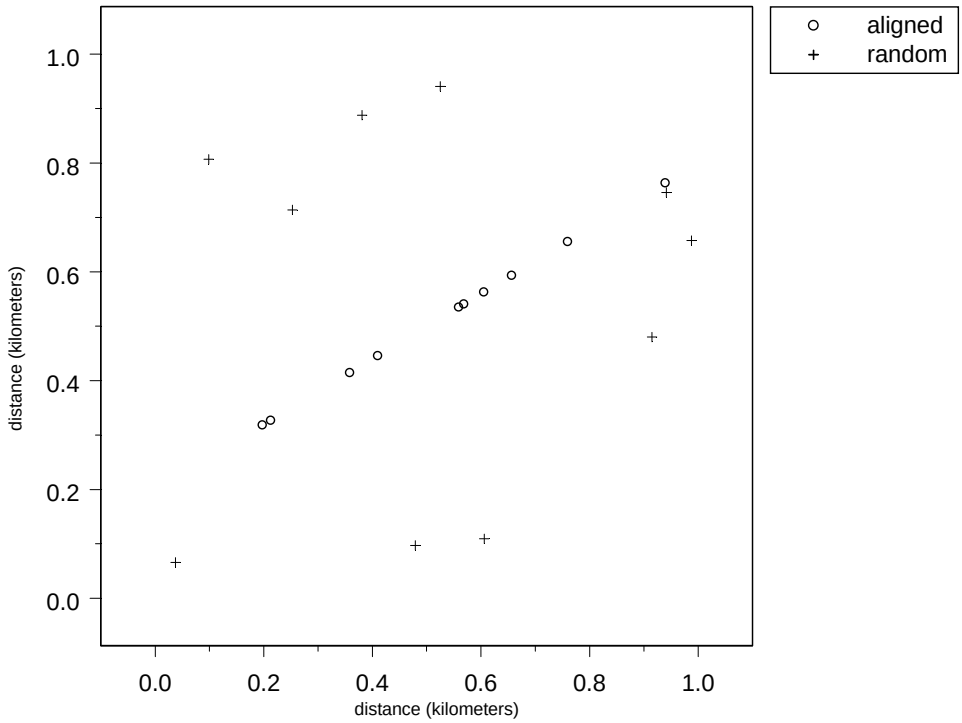
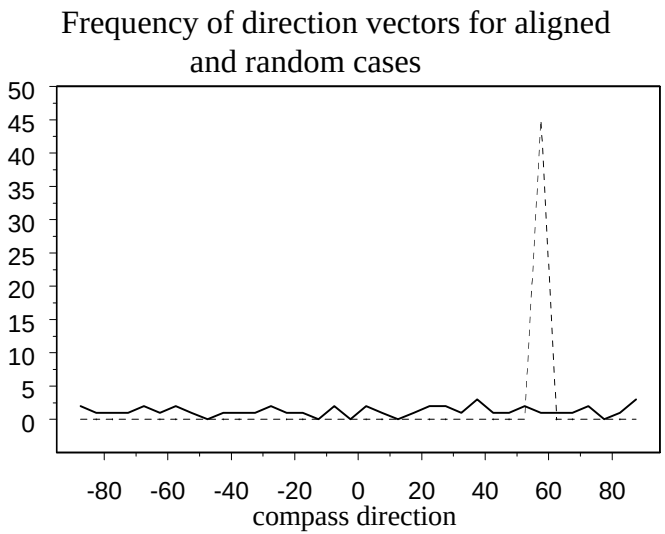
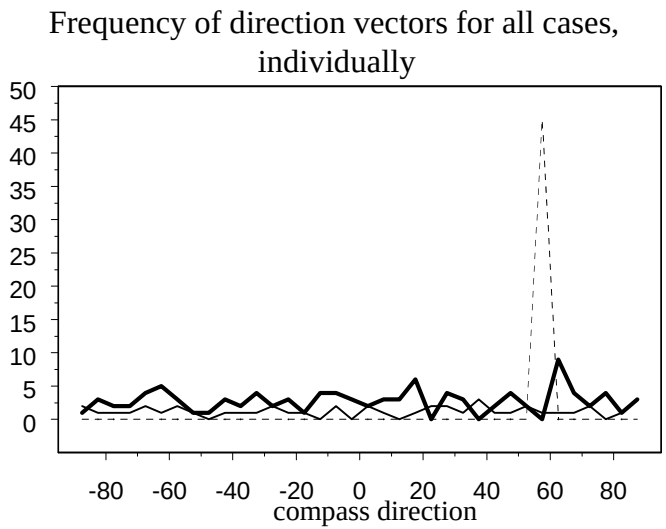
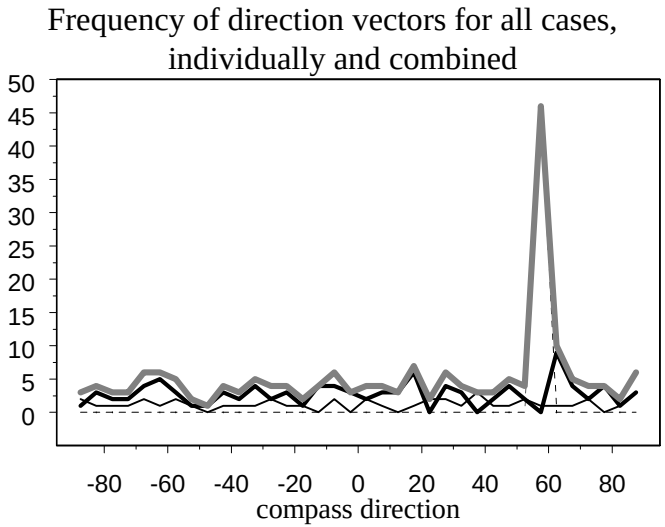


Figure 26. Location of 10 randomly placed points and ten points placed randomly along a line according to y=0.6x+0.2 (an arbitrary line). The direction vectors connecting the random points will not show a preferential direction; the direction vectors connecting the points along the line will.

The direction vectors for this point set will result in a total of 190 direction vectors (20*19/2), 45 of which are generated by pairings of aligned points (10*9/2), 45 of which are generated by pairings of random points, and 100 of which are generated by pairing a random point with an aligned point. Figure 27 (top left) is equivalent to figure 20, and







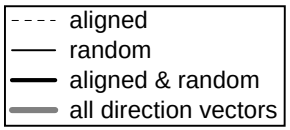


Figure 27. The direction vectors for a set of combined aligned and random points are shown here broken down into components: aligned-aligned and random-random direction vectors (upper left), then adding the aligned-random direction vectors (upper right), then all components with the total frequency of direction vectors (bottom left). Note how the aligned-random direction vectors show a random pattern.

figure 27 (top right) adds the direction vectors for the random-aligned pairings. The direction vectors in figure 27 (top right) add together to create the distribution of direction vectors for the whole point set, shown in Figure 27 (bottom left). Note that the distribution of direction vectors for the random-aligned pairings is roughly uniform. One way to think of this result is that a single point, when added to a random point set, will produce direction vectors that appear random. The direction vectors for the random-aligned pairings is the combined direction vectors for each aligned point with the random point set.

One can consider the direction vectors associated with the pairing of the aligned points as a *signal*, and the rest of the direction vectors as *noise*. This can be defined by a noise factor, *k*, where n_a is the number of aligned points, and therefore

1) $kn_a = n$

is the total number of points in the point set. In general, increasing *n* will cause percentage of signal in the direction vector distribution to decrease towards

2) $\frac{1}{k^2}$.

so as the size of the point set increases, the proportion of direction vectors considered *noise* approaches $1-\frac{1}{k^2}$ (see Appendix I for proof). If the signal-to-noise ratio (in this case the ratio of the direction vectors considered signal to the total number of direction vectors) can be estimated from the direction vector distribution, n_a can be estimated by the relationship defined above:

$k = \sqrt{\frac{1}{R}}$, where $R = \frac{n_a(n_a-1)}{kn_a(kn_a-1)}$

and by *1)*, $n_a = \frac{n}{k} =$

3) $\frac{n}{\sqrt{\frac{1}{R}}}$

where R is the proportion of direction vectors considered signal to the total number of direction vectors, n_a is the number of aligned points, *n* is the total number of points in the set, and *k* is a constant.

For large values of *n* the equation simplifies further, since

$$R = \frac{n_a(n_a-1)}{kn_a(kn_a-1)} \approx \frac{n_a(n_a-1)}{n^2},$$

then for large values of *n*

$$4) \qquad \frac{n}{\sqrt{\frac{1}{R}}} \approx \frac{n}{\sqrt{\frac{n^2}{n_a(n_a-1)}}} = \sqrt{n_a(n_a-1)}$$

For example, consider the direction vector distribution for the point set above, as if the structure of the point set were unknown. The distribution suggests an alignment of points at north 55º east, and randomly distributed direction vectors otherwise. The mean frequency of the data, excluding the peak direction of 55º east, is 4.11 direction vectors per 5º. Assuming this applies to all directions, there are 148.11, or 148, direction vectors that are in the noise category, and therefore 42 direction vectors in the signal category. The ratio of direction vectors considered signal to the total number of direction vectors is therefore 42/190. By equation 3, the number of aligned points in the point set is $20/\sqrt{190/42}$ =9.4, or 9 points. Considering the direction vector distribution for each individual point, the points with their direction vector mode at 55º east are logical candidates to be the aligned points. In this synthetic example there are 10 points with equal frequency of direction vectors for 55º east, with no clear way of distinguishing them. The model considers only the average random response, and has not accounted for the individual character of this dataset. In this synthetic example the number of aligned points is under-predicted because the model assumes that there is a random component to all directions. In this small sample, no random point fell in the aligned direction, and thus the random signal was absent in the aligned direction. Since the calculated number of aligned points is approximate, bringing additional information to bear on borderline cases would be the next step in any investigation.

Clusters of points can also produce a peak in the direction vector distribution. Figure 28 shows the distribution resulting from taking the 10 random points in a 1 km by 1 km area described above, and adding one point to the point data set at ten times the range (10 km away). In this modified dataset the random points behave like a cluster, and the added point is an outlier. As can be seen from the figure, the direction vector causes a peak in the distribution of direction vectors. Two clusters sufficiently distant from one another essentially form a line between each other, and therefore cause a peak similar to collinear points. As with the lineaments, clusters of points can the bias the results of the analytical method discussed in the previous paragraph.

The previous example has shown that the direction vector distribution for a point set can be thought of as different components adding together to create the total distribution. Given a direction vector distribution, one can construct simple models of linear structure and attempt to separate the components of the direction vector distribution. For a simple model of aligned points and random points, one can estimate the number of aligned points, and, using the direction vector distribution for each individual point, develop a set of candidate aligned points. In this case the direction vector distribution can be used as a filtering process.

For the observed damage dataset, the declustered data is used to avoid the problem the large cluster in the town of Los Gatos will cause. Small clusters in the data will still have an effect, but the largest effect will be minimized. The histogram of the distribution of direction vectors for the declustered dataset is shown in figure 29 (small-dashed line; in this figure, as in figure 20, the distribution has been quantized into 5º bins

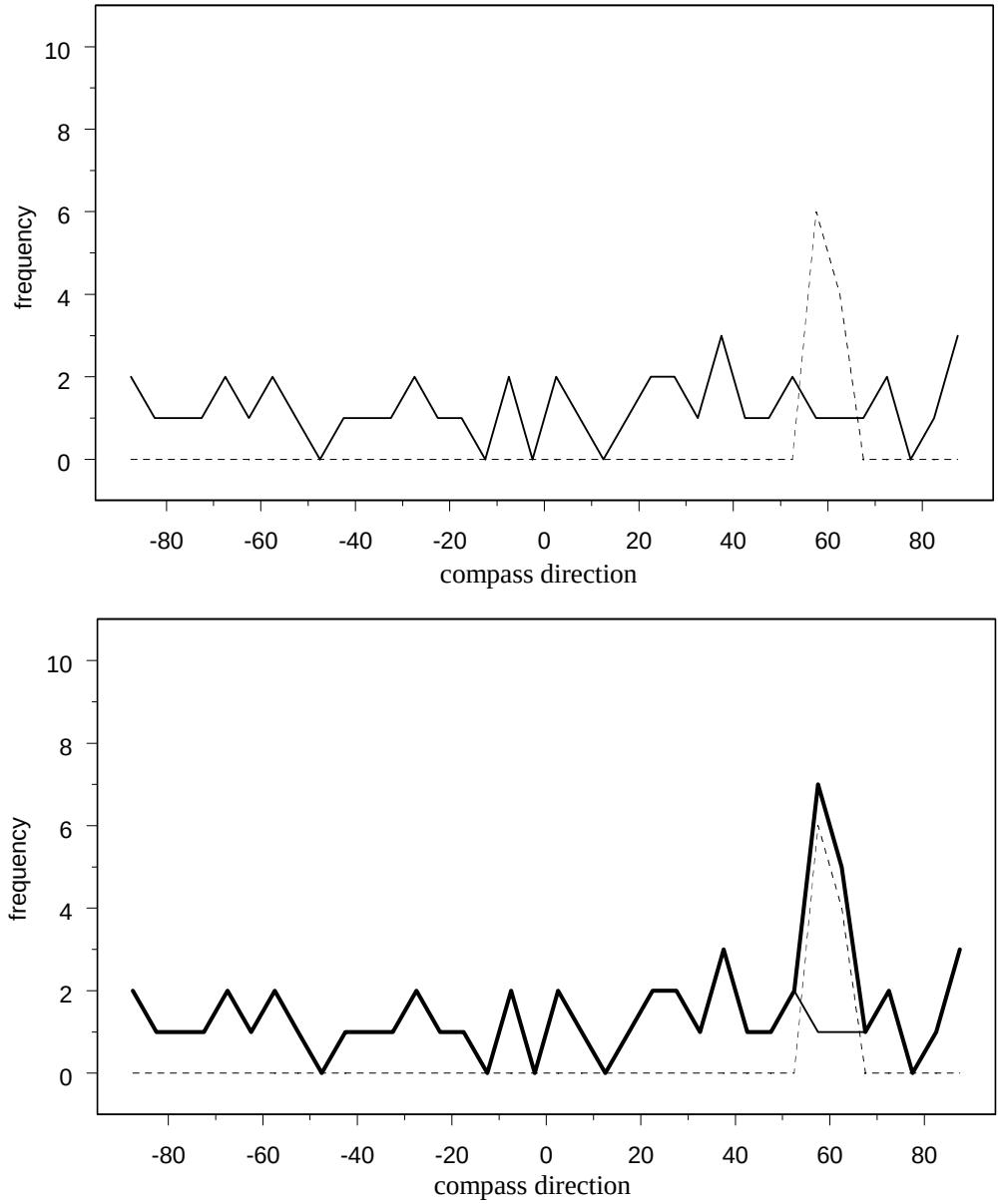


Figure 28. Distribution of direction vectors for the set of ten randomly located points and an additional outlier point located at ten times the range of the random points. The distribution of direction vectors for the random set (thin solid line) is compared with that of the outlier-random direction vectors (dashed line) in the top figure, and both are compared with the direction vector distribution for the entire dataset (thick solid line) in the bottom figure.

and the bin mid-points connected by lines instead of using histogram bars, in order to emphasize the shape of the distribution and to be able to subtract one distribution from the other). It has a similar shape to the complete observed damage dataset. The largest number of random direction vectors possible, without exceeding the observed number of

direction vectors at any given point, is fit to the distribution of direction vectors for the declustered dataset (figure 29, large-dashed line), and subtracted from it, to obtain the residual distribution (figure 29, solid line). The residuals have a clear peak from 55º to 40º northwest.

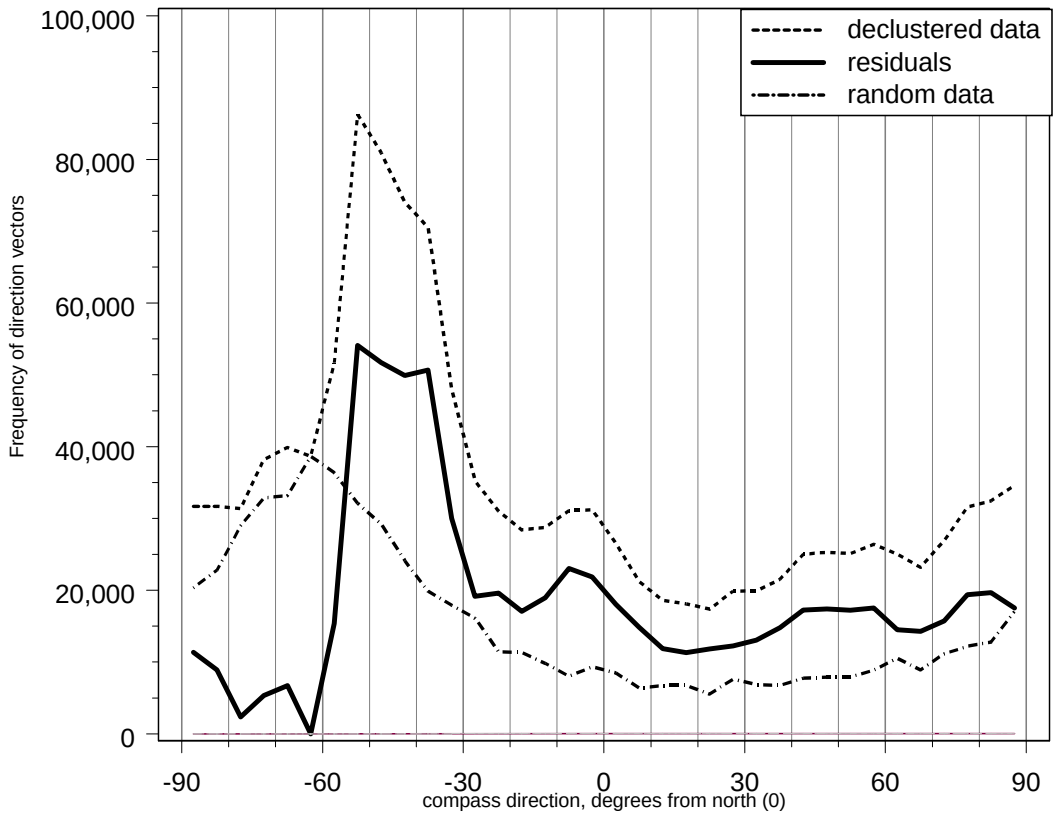


Figure 29. Distribution of direction vectors for the declustered damage dataset, a random dataset, and the residuals of the distribution of the declustered data minus the distribution from the random data. The largest subset possible of direction vectors forming the random distribution was fit to the declustered distribution.

Figure 30 shows the results of choosing the damage points with their individual direction vector mode between 55º to 40º northwest. That is, for any point P_i, consider all direction vectors for that point: $\{(P_i,P_1),(P_i,P_2),(P_i,P_3),...,(P_i,P_j),...,(P_i,P_n)\}$, excluding (P_i,P_i), and select the mode of this distribution as the mode of the direction vectors for the point P_i. Figure 30 considers points whose mode is between 55º to 40º northwest. When separated into groups according to their mode, the alignment of points is moderately linear, strongly linear, weakly linear and somewhat clustered. The 157 points having a direction vector distribution mode of 50º northwest, shown in the upper right of figure 30, show the strongest linear pattern. There are 51,751 residual direction vectors for this direction. Estimating the expected number of points from the direction vectors using a model of linear and random points leads to $n_a \approx \sqrt{51751} \approx 227$,

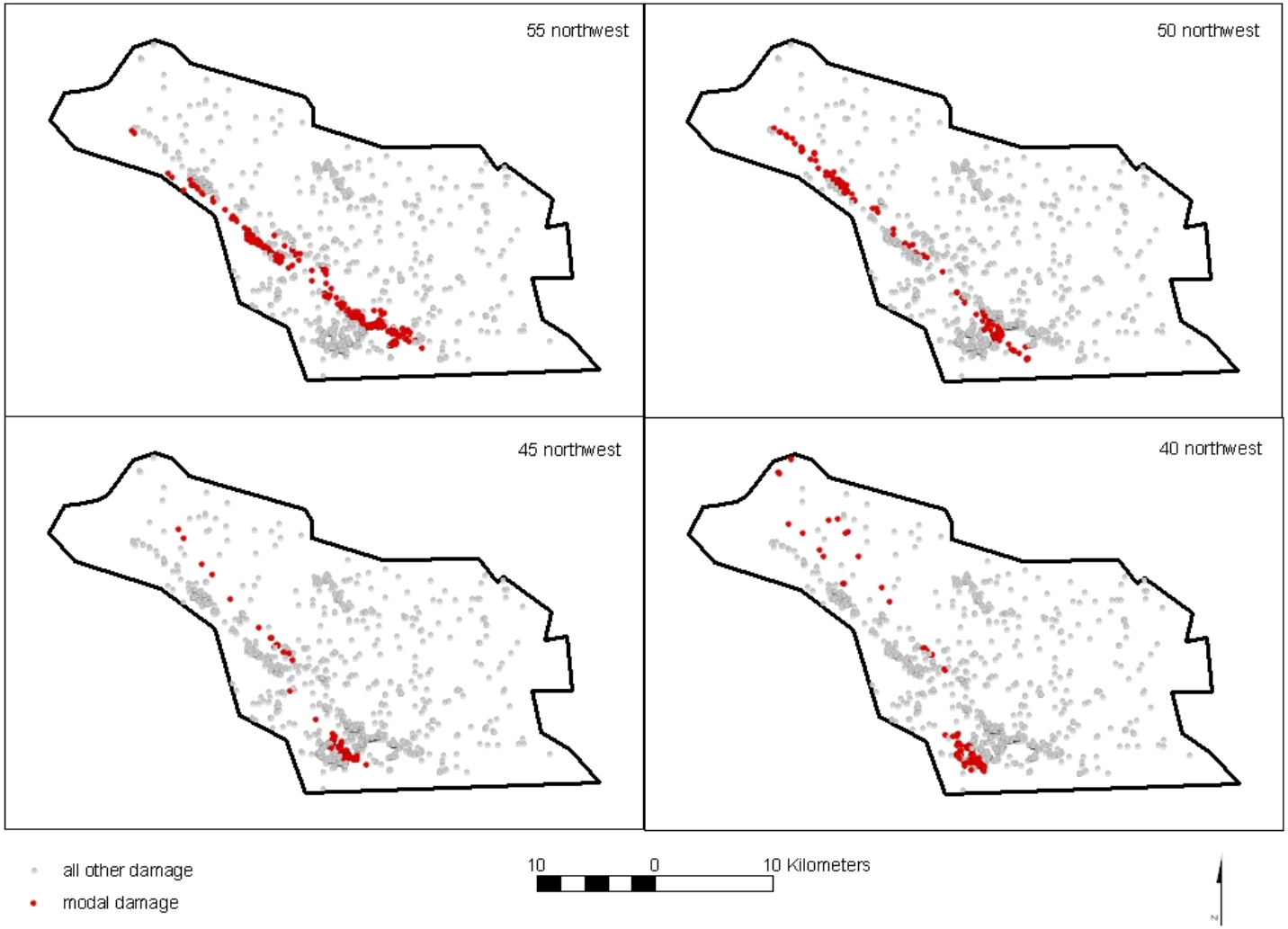


Figure 30. Declustered dataset showing observed damage points with the mode of their individual direction vector distribution at 55, 50, 45, and 40 degrees northwest (modal damage). Note that 50 degrees northwest shows a confined linear pattern, and 40 degrees northwest shows a clustered pattern.

overpredicting the observed 157 points having a direction vector distribution mode of 50º by about 30%. If one considers the entire peak from 55º to 40º northwest, the direction vectors must be added together in equation 4. The estimated number of direction vectors for a peak signal from 55º to 40º northwest is 206377, out of a total of 1246572 direction vectors in the entire distribution. This results in an estimate of $n_a \approx \sqrt{206377} \approx 454$. The total number of damage points with a mode from 55º to 40º northwest is 741. The discrepancy between observed and predicted aligned damage is probably because the damage is not perfectly aligned, as the model demands, but are more dispersed, which causes fewer direction vectors to be aligned in one direction. It therefore takes more points to generate the same number of direction vectors in the aligned direction. It also indicates that the model of a linear alignment of points plus a random set of points is not completely applicable to the declustered damage dataset. This is most apparent for the points with a direction vector distribution mode of 40º northwest, where outlier points in the northern portion of the study area can be seen responding to clusters of points in the southern portion. The declustering algorithm reduces, but does not eliminate, this effect. From the declustered data a clear linear trend emerges for 50º northwest, and a reasonably linear trend for 55º northwest. For the directions 45º northwest to 40º northwest the trend is less linear and includes some signs of clustering patterns.

The linear trend of 50º northwest apparent within the damage data can be compared with the mean trend of the mapped lineaments. The lineaments are aligned in a northwest-trending direction (figure 3). If each straight segment of the mapped lineaments is considered as a vector (so that a single mapped lineament that consisted of many segments would be considered many vectors, connected head-to-tail), the mean direction of the lineaments can be found by calculating the mean direction of all of the vectors. The mean direction of the vectors can be found by adding the vectors and calculating the direction of the resulting sum. The component vector information is easily found by extracting the coordinate information from the GIS (appendix 2). The mean direction for the mapped lineaments is 49º northwest. Thus the prominent linear pattern within the damage data matches quite closely with the mean direction of the mapped lineaments, again suggesting a common factor, faulting along the range front of the Santa Cruz Mountains, links them.

Problems with the model

The method of fitting a model of aligned points plus random points to a scattered point dataset considers points aligning in a single direction. For practical application to faulting in geology, this may be somewhat limited, because faults may bend, or show more than one preferential direction. For investigation of points following a curving line, or for conjugate sets, the model for this simple analysis breaks down. In this particular area it is fortunate that nature seems to have fairly regular behavior.

The method is also affected by the width of fault zones. A fault zone adds short-distance noise to the distribution of direction vectors, and the resulting direction vector peak in the direction vector distribution is not as pronounced. This leads to an underprediction in the number of points producing the “signal”.

Lastly, other structures in the scattered point dataset, such as clusters and outliers, or multiple clusters, can interfere with the signal of aligned points. These structures are neither linear nor necessarily random, and must be investigated separately.

SUMMARY AND CONCLUSIONS

Three analytical techniques relevant to the analysis of point patterns were discussed: (1) testing whether or not points are closer to some object than would be the case with complete spatial randomness, (2) testing whether or not a set of points exhibited complete spatial randomness, and (3) a method of filtering a point set for a subset of points that are aligned. The techniques were demonstrated using a dataset of damage caused by the Loma Prieta earthquake and a dataset of mapped lineaments within the damaged area.

The first model of random behavior defined the relationship of points to other objects (point, line, or polygon features), applying techniques previously developed in the field of urban planning (Okabe and Fujii, 1984). By applying basic probability theory, the theoretical distribution of the distance a set of randomly located points to the objects, for an arbitrarily shaped study area, can be constructed and used as a null hypotheses of randomness for comparison with observed data. The first model of random behavior defined the relationship of random points to a set of linear features, in this case lineaments.

Since the above analysis can be sensitive to inhomogeneities in the dataset, such as clustering, a declustering process was proposed that models a cluster as a circular phenomenon whose point density decays from the center of the circle according to a gaussian function. When applied to the damage data, the declustering diminished the bias present in the damage-lineament analysis. The results demonstrate that the cluster is the most significant pattern in the data in terms of damage concentration and that the cluster significantly influenced the results of the initial statistical analysis.

The second model of random behavior defined the relationship of points to each other. The theoretical random distribution of direction vector orientations, defined by each and every pairing of points in a set of points, was constructed using numerical methods. This method of developing a random distribution of direction vectors was compared with that developed using Monte Carlo methods, and the two were found to agree quite closely.

The method outlined above was used to demonstrate that the locations of the damage data are not randomly located throughout the study area. The distributions were compared using a Chi-square test based on circular ranks, and the null hypothesis of randomness was rejected. The preferential alignment of the random process was found to be about 60º northwest, whereas that of the observed process was 40º northwest. This indicates that the observed process is biased to the north. Furthermore, the random distribution of direction vectors is unimodal, whereas the observed distribution of direction vectors is multimodal.

The distribution of direction vector orientations for the damage data suggested that a theoretical model, composed of the union of a set of points aligned along one direction and a set of randomly located points, was a reasonable model for the patterns observed in damage data. Using this as a base model, it was discovered that the cluster process dominates the signal of the entire dataset and that a linear signal can be seen when analyzing the declustered dataset, though some of the signal from clustering in the data is still present. The most significant linear trend is 50º northwest, where the model

predicted 227 points would be aligned in that direction, and in the observed damage dataset there were 157. The linear trend in the damage data of 50º northwest matches quite closely the mean trend of the mapped lineaments, which average 49º northwest.

Numerical methods can be used to develop null models of complete spatial randomness over (planar) irregularly shaped study areas. These models can be generated quickly using modern GIS software. They are applicable to determining if a set of points is associated with some other set of objects (points, lines, or polygons) or determining if a set of points itself exhibits complete spatial randomness. The direction vectors resulting from a point dataset can be used to filter a point dataset to find points that are aligned. This filtering process can be used to separate the point dataset into component patterns of aligned and non-aligned points, provided a single linear pattern exists and clustering of points is not excessive.

ACKNOWLEDGMENTS

The following people generously provided data and time for thoughtful discussions: Carl Wentworth, Kevin Schmidt, and Robert Jachens. Reviews by Donald Singer and Robert Simpson were very helpful. Special thanks to Michael Barall, who greatly improved the mathematical underpinnings of the methods. This work would not have been possible without their help.

REFERENCES

Brabb, E.E., Graymer, R.W., and Jones, D.L., 2000, Geologic map and map database of the Palo Alto 30' x 60' quadrangle, California: U.S. Geological Survey Miscellaneous Field Studies Map MF–2332, 32 p.

Connover, W.J., 2003, Practical non parametric statistics: John Wiley & Sons Inc., 584 p.

Diggle, P.J., 2003, Statistical analysis of spatial point patterns: Arnold, London, 159 pp.

Fisher, N.I., 1995, Statistical analysis of circular data: Cambrdge University Press, Cambridge.

Graymer, R.W., Bryant, W., McCabe, C.A., Hecker, S., and Prentice, C.S., 2006, Map of Quaternary-active faults in the San Francisco Bay region: U.S. Geological Survey Scientific Investigations Map 2919, map scale 1:275,000.

Hitchcock, C.S., Kelson, K.I., and Thompson, S.C., 1994, Geomorphic investigations of deformation along the northeastern margin of the Santa Cruz Mountains: U.S. Geological Survey Open File Report 94-187, map scale 1:24,000.

Hitchcock, C.S., and Kelson, K.I., 1999, Growth of late Quaternary folds in southwest Santa Clara Valley, San Francisco Bay area, California: Implications of triggered slip for seismic hazard and earthquake recurrence: Geology, v. 27, no. 5, p.391–394.

Kaluzny, S.P., Vega, S.C., Cardoso, T.P., and Shelly, A., 1996, S+SpatialStats: User's manual for Windows and UNIX: Springer-Verlag, New York, 327 pp.

Langenheim, V.E., Schmidt, K.M., and Jachens, R.C., 1997, Coseismic deformation during the 1989 Loma Prieta earthquake and range-front thrusting along the southwestern margin of the Santa Clara Valley, California: Geology, v. 27, p. 387-390.

McLaughlin, R.J., Clark, J.C., Brabb, E.E., Helley, E.J., and Colón, C.J., 2001, Geologic Maps and structure sections of the southwestern Santa Clara Valley and southern Santa Cruz Mountains, Santa Clara and Santa Cruz Counties, California: U.S. Geological Survey Miscellaneous Field Studies Map MF–2373, 13pp.

Okabe, A., and Fujii, A., 1984, The statistical analysis through a computational method of a distribution of points in relation to its surrounding network: Environment and Planning A, v.16, p.107-114.

Okabe, A., Fujii, A., Oikawa, K., and Yoshikawa, T., 1988, The statistical analysis of a distribution of activity points in relation to surface-like elements: Environment and Planning A, v.20, p.609-620.

Rock, N.M.S., 1988, Numerical geology: Springer-Verlag, Berlin.

Schmidt, Kevin M., Ellen, Stephen D., Haugerud, Ralph A., Peterson, David M., and Phelps, Geoffrey A., 1995, Breaks in pavement and pipes as indicators of range-front faulting resulting from the 1989 Loma Prieta earthquake near the southwestern margin of the Santa Clara Valley, California: U.S. Geological Survey Open-File Report 95-820, 33pp.

Upton, G., and Fingleton, B., 1985, Spatial data analysis by example, Volume 1:Point pattern and quantitative data: Wiley, Chichester, England.

U.S. Geological Survey, 2007, October 17, 1989 Loma Prieta earthquake http://earthquake.usgs.gov/regional/nca/1989/

Wentworth, Carl M., Blake, M. Clark, Jr., McLaughlin, Robert J., and Graymer, Russell W., 1998, Preliminary geologic map of the San Jose 30 x 60 minute quadrangle, California: A digital database: U.S. Geological Survey Open-File Report 98-795, 52p.

APPENDIX I

Given: the union of two independent sets of points defined on a plane, one a set of randomly located points, the other a set of points that fall along an arbitrary line, such that the total number of points in the union of the two point sets is n. Direction vectors are generated by the pairing of points, with one direction vector for each point-pair. The complete set of direction vectors is generated by the unique pairing of all points.

Prove: the ratio of the number of direction vectors from the aligned point set to the number of direction vectors from the union of the two point sets approaches $\frac{1}{k^2}$ as n approaches infinity.

Let $n_a =$ the number of points in the aligned set

$kn_a = n =$ the number of total points in the union of the two sets

$n_a(n_a - 1)/2 =$ the number of direction vectors in the aligned point set, here considered *signal*

$kn_a(kn_a - 1)/2 =$ the total number of direction vectors in the union of the two point sets

Proof: the ratio of direction vectors considered *signal* to the total number of direction vectors for a given point set is

$$\left(\frac{\frac{n_a(n_a-1)}{2}}{\frac{kn_a(kn_a-1)}{2}} \right)$$

$$= \frac{n_a(n_a-1)}{kn_a(kn_a-1)}$$

$$= \frac{(n_a-1)}{k(kn_a-1)}$$

The limit of this ratio as $n \to \infty$ is

$$\lim_{n\to\infty}\left[\frac{(n_a-1)}{k(kn_a-1)}\right] = \frac{1}{k}\lim_{n\to\infty}\left[\frac{(n_a-1)}{(kn_a-1)}\right] = \frac{1}{k^2}\lim_{n\to\infty}\left[\frac{(n_a-1)}{\left(n_a-\frac{1}{k}\right)}\right]$$

$$= \frac{1}{k^2}$$

APPENDIX II
Technical specifications for various spatial analysis tasks

Declustering point data

1. Convert the (vector) point pattern to a continuous surface by generating a density map.
2. Digitize a cross-section line across the anomaly (visible on the density map); add vertices to the line at regular sampling intervals.
3. Convert the cross-section-line vertices to points; extract the value of the density map at the location of the points.
4. Output these sampled values (Y values) and their respective distance along the cross-section-line (X values) to a statistical package and model the anomaly.

Approximating the probability distribution of the distance to the nearest line for complete spatial randomness

1. If the study area is in vector format, convert the study area to raster in order to generate an approximation of a space-filling set of points.
2. Convert the raster (cells) to (vector) points.
3. Find the distance to the nearest line for each point, which as a set represent the entire study area (e.g. the NEAR command in Arc/Info). Ensure that each record includes the line id number and distance. This will allow for further analysis using subsets of data based on the individual lines.
4. Output the distance data to a statistical package. This distribution of distance is an approximation of the probability distribution of the distance to the nearest line for complete spatial randomness within the study area.

Finding the observed probability distribution of the distance to the nearest line for a set of points

1. For the observed data, find the distance of each observation to the nearest line (e.g. the NEAR command in Arc/Info). Ensure that each record includes the line id number, point id number, and distance. This will allow for further analysis using subsets of data based on the individual lines.
2. Output the distance data to a statistical package. This is the observed distribution of the distance of points to lines, and can now be compared with the approximation of the theoretical distribution.

Finding the average orientation of lines based on individual line segments

1. Consider each pair of vertices along the length of a line as a vector.
2. Extract the coordinates of each vertex, in order, along the line, and calculate $\Delta x = (x_2 - x_1)$ and $\Delta y = (y_2 - y_1)$ for each pair.
3. To add the vectors and find the direction of the average vector, calculate $\tan^{-1}\left(\frac{\sum \Delta x}{\sum \Delta y}\right)$ and convert the result to the compass direction.