Commentary: Introductory Comments to Some Applied Papers by David R. Brillinger, by Tore Schweder and Haiganoush Preisler

In addition to statistics, David took care in developing my attitude as a scientist - and he wrote a poem of his own in my draft thesis about whales and statistics. He also cared for us personally. We were invited to use Lorie's and David's house when they went to New Zealand in the summer of 1973. Our newborn child spent her first time out of Alta Bates hospital in their house. David also gave me support in a more touchy matter. I was on a US Navy grant, and felt uneasy when I realized that I had to acknowledge the grant in a publication. Strike it in the last galley, was David's advice – which I in the end did not follow. And there was fun, also outside the soccer field. David suggested the movie "The harder they come". My son, an aspiring reggae musician, was happy to find the Jimmy Cliff LP in my old stock. In the last couple of years we have been lucky to have David as an advisor in our Centre for Ecological and Evolutionary Synthesis in Oslo, and to have David repeatedly visiting.

Empirical modelling of population time series data: The case of age and density dependent vital rates [1980]

A stochastic matrix model is used to study a population of sheep blowfly observed over two years in a lab. The flies were kept in a cage, and fed on a constant diet. The population experienced substantial fluctuations in size over the observational period. Matrix models for stage-structured populations like the sheep blowfly have become popular (Caswell (2000) *Matrix Population Models: Construction, Analysis, and Interpretation*, is cited some 1900 times).

The hypothesis behind the experiment was that competition for resources would occur only in egg laying and that the population fluctuations were due to variability in recruitment. Mortality was thought only to depend

P. Guttorp and D. Brillinger (eds.), *Selected Works of David Brillinger*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4614-1344-8_18,

on age. By fitting a product model to age-and abundance specific survival probabilities by weighted least squares, it is found that age-specific mortality does depend on abundance, and also on abundance two days earlier. By residual analysis it is also found that there are further dependencies.

The present paper is an early study with vital parameters, particularly the mortality rate, depending on population size and being affected by random variation. It has influenced the field of population dynamics to make more use of rather standard statistical methodology.

Learning a Potential Function from a Trajectory [2007] This short Signal Processing Letter presents stochastic differential equation models for moving objects where the drift term is the negative gradient of a potential function, providing a formal background and a general discussion missing in the literature. These models can have regions of attraction, absorbtion or repulsion. Various interesting and useful potential functions that are linear combinations of given differential functions are considered. Being linear in unknown parameters, they can be estimated by linear regression although with stochastic regressors. Asymptotic theory is presented for the potential function estimator based on standard assumptions on conditional independence and zero mean residuals. A similar regression model is used in the soccer study (Brillinger 2007b) discussed below. For curved potential functions one might wonder whether the residual terms really have zero mean. Another question is how the asymptotics of the estimated potential depends on the potential function itself. If the attraction, say to a point, is sufficiently strong, a single trajectory might for example not provide enough information to allow first order convergence to normality.

A Potential Function Approach to the Flow of Play in Soccer [2007] Soccer, or football as we say outside USA and Canada where this game is by far the most popular of sports, is a game played by two teams of 11 players each. The play field is rectangular about 105 by 68 m, with a goal at each short end. The purpose is for each team to have the ball inside the goal of the other team as many times as possible, and have the ball inside its own goal as few times as possible. The ball is passed from player to player within a team, usually by the kick of a foot, until a goal is scored, the ball gets off the field or is picked up by the other team or is lost because an arm or hand is used or some other rule is broken.

David got interested in soccer during his years at LSE. In the early 1970s we were several Norwegian students at the Berkeley department. We got a Norwegian newspaper to the coffee room, and David was quick to grab it to get news of soccer in England.

The purpose of this article is to establish a statistical framework for describing and simulating how a game of soccer develops. This is done by breaking the game out in spells of ball occupancy. A spell is a succession of passes of the ball within a team. An unusually long spell in the 2006 World Cup game between Argentina and Serbia-Montenegro is studied in detail. It had 25 passes and ended with a goal for Argentina. The spell is characterized by the position r_i of the ball when pass i is initiated, and also the time t_i. A potential function H that describes the spatial succession $r_i \to r_{i+1}$ for given times is assumed:

$$(r_{i+1} - r_i) / (t_{i+1} - t_i) = -\nabla H(r_i) + noise.$$

The potential function is assumed linear in some parameters, and so is the gradient. These parameters are estimated by least squares from the positions and times of the spell. The estimated potential function is a bit skewed towards the left of the field seen from the Argentinian side, and might be symmetrized when used to simulate a game. To simulate a game, a potential function is also needed for the other team, and also a way to simulate time points within spells and a stopping time for spells.

Since passes generally are made towards the opposing goal where the potential function is steepest, the least squares approach taken in the paper will bias the estimated potential function towards less steepness. A partial fix is to also include the curvature (the Hessian) in the regression. Another point to be made is that simulating soccer games are done in different computer games. How are these simulations carried out relative to the method proposed here?

Analyzing games of soccer, i.e. pulling them apart in their basic elements like spells, and characterizing teams by their estimated potential functions and other characteristics might be useful for understanding games and for the training of a team. One point of particular interest might be to identify to what degree and in what aspects game results are determined by simply combining the features of each team without further interaction than independent succession of spells, and in what aspects more complex interaction must be invoked. If no such further interactions are of importance, a team should be trained without regards to qualities of the next opposing team.

The use of potential functions in modelling animal movement [2001] Potential functions are used in the physical sciences to model the motion of planets or particles in a field of gravitation or other forces determining velocity and direction. A two-dimensional stochastic differential equation (SDE) provides a stochastic version of the deterministic potential function model. The drift term in the SDE is then the negative gradient of the potential function. The authors use an SDE to model the motion of elk (and mule deer) in an enclosed forest. The extensive data was obtained from researchers who fitted a number of animals with collars containing Loran -C receivers. The position of a tagged elk is recorded about every minute, but with a measurement error of some 50 meters. Under the SDE model, the assumed potential function is estimated from the spatial distribution of elk positions, assuming this distribution to be the stationary distribution of the SDE. An interesting question is whether the function estimated this way by a kernel method really could be an estimate of a proper potential function. This is tested somewhat informally by the Student statistic found by comparing the two cross differentials of the estimated function. The authors find that the existence of a potential function cannot be rejected. Gray scale graphs depict the estimated potential function showing that elks tend to be in the northern end of the area during day, while more to the south during night.

Do individual elks move about according to a Markov process, say an SDE, and independently of each other? The potential function model assumes this. Although these basic assumptions cannot be tested within the SDE model, the model is very useful in summarizing position data for tagged animals to elucidate questions about habitat selection and foraging behavior, and also the effects of vehicular traffic and fences on animal behavior.

The paper shows how the SDE model might be used for animal motion data, and it finds its place in the sequence of related papers on models and numerical methods that David has together with biologists and fellow statisticians.

Elephant-seal movements: Modelling migration [1998] Elephant seals migrate between their rookeries at the California coast and their feeding areas in the North Pacific close to the Aleutian islands twice a year. Data from individuals fitted with tags measuring light conditions by time of the day when surfacing to breath, in addition to other variables were considered. The data allow the migration route to be tracked by estimating the position

each morning by the length of daylight and the time of sunrise. The seals are essentially following a great circle between foraging area and haul-out. How they manage to navigate as precisely as they do is not known, but the authors speculate that the seals utilize the global magnetic field. In order to develop testable hypotheses for seal migration, formal models are developed. These models are cast in the form of stochastic differential equations for diffusion on the surface of the globe. They are of the Ornstein-Uhlenbeck type. From Brillinger (1997), the basic model for the case of steady drift along a meridian is set down:

$$d\theta_t = \left(\frac{\sigma^2}{2\tan(\theta_t)} - \delta \right) dt + \sigma dU_t, \quad d\phi_t = \frac{\sigma}{\sin(\theta_t)} dV_t$$

for latitude θ and longitude ϕ and standard Brownian motions U and V, all measured in radians. In this formulation the speed δ is negative when the motion is towards the North Pole. An alternative model is discussed below. If the target and the point of departure are at different meridians, the equations for the diffusion along a great circle are obtained by a trigonometric transformation of the above equations.

In this model, including measurement errors in the daily positions, approximate maximum likelihood estimators are obtained. Based on data from a seal migrating from the foraging area towards its rookery, estimates are obtained for speed along the great circle δ and for the diffusion standard deviation σ. The point of departure for this seal was estimated as the mean position over the days thought to be spent on foraging ahead of the migration towards the California rookery. The target, the seal's rookery, is a precise point, but the point of departure needs perhaps not be a well defined point. It might be preferable to model the foraging area as an area and not a point, say by a bivariate normal distribution located at the center of the foraging area.

In case navigation is done continuously by the magnetic field, continuous time modelling as above is appropriate. If however navigation is done celestially, discrete time modelling might be preferable. In that case navigation is prevented during daylight when the stars are invisible, and also those nights with clouds on the sky.

Tag technology has developed considerably since the elephant seal data were obtained. Modern tags would measure position by GPS at each surfacing, with very little measurement error. Such tags are widely used to

understand animal behavior, and diffusion models for migration on a sphere like the one above should be in considerable demand.

An alternative model for an Ornstein-Uhlenbeck process on the sphere with drift towards the North Pole and with attraction to a meridian at longitude m is $d\theta_t = \delta dt + \sigma dU_t$, $d\phi_t = \gamma(m - \phi_t) + dV_t$. Here θ_t is latitude and ϕ_t is distance from (θ_t, m) along a great circle perpendicular to the meridian, and dU is random latitudinal disturbance and dV is the same along the perpendicular great circle. In this model, θ is causally independent of ϕ, and is therefore a one-dimensional linear SDE process. The exact likelihood for observations at discrete points in time is available. The expected velocity along the great circle is δ and the push of reverting back to the meridian at m along a perpendicular great circle at position ϕ_t is γ. This model is, perhaps, a more transparent model than that in the paper.

Random process methods and environmental data: the 1996 Hunter lecture [1997] Environmental processes like weather, river flow, earth quake damage etc. are essentially dynamic and nearly always affected by random variation, and random processes are fundamental in modelling them. This is the basic message of the paper, and three environmental processes are considered to illustrate the use of stochastic process models and statistical inference in environmental science.

In the first analysis nearly a century of daily river height of Rio Negro at Manaus, Brazil are analyzed. The question is whether there is an increasing trend in the flow of water out of the Amazon basin. There is considerable seasonal variation in river flow, and the approach is to fit a model with a trend plus a seasonal component and one for daily variability. Rather than modelling these components parametrically, year specific seasonal component are estimated by the median annual curve. The trend is only assumed to be non-decreasing and is estimated non-parametrically from the seasonally adjusted series. The estimated trend curve is by construction non-decreasing, and to judge whether it estimates a truly increasing trend, data were simulated from the model assuming no trend. The simulated data were analyzed in the same way as the observed data, and from visual comparison, the conclusion is that there is a soupçon (touch of) significance. This ingenious non-parametric time series analysis could be generalized. One might ask whether it yield other conclusion than more traditional parametric analyzes?

Damage due to the Loma Prieta earthquake with epicenter close to Santa

Cruz, California, is the theme of the second analysis. At various localities in the greater Bay area the earth quake damage was measured on an ordinal scale with 12 levels of increasing severity. The purpose of the analysis is to extrapolate these damage measurements to the whole of the affected area. The degree of damage at an affected locality at (x, y) is modelled as a multinomial based on a smooth spatial function $g(x, y)$ plus an extreme value distributed random variable. The contours of the estimated damage function g are shown. Perhaps the predicted value of the ordinal damage score might have been be more interesting. Other distributions than the extreme value could actually have been chosen. The predicted damage score is broadly invariant to the choice of distribution, while g has the scale of the chosen distribution.

In the third analysis the problem is to estimate the average velocity at which weather moves from west to east on the Globe. The data consist of 500 millibar pressure fields across the surface of the earth over a five day period, with two measurements (pressure fields) per day. The pressure field is modelled as $Y(x,t) = g(x) + h(x - vt) + noise$, where x is longitude in radians, v is velocity in radian/hour and t is time in hours. The longitude specific component g cancels in the difference $Y(x, t+1) - Y(x,t)$ and these differences are used to estimate the velocity by least squares.

David describes statistics as the science of using data wisely. He further makes the basic general remark that for problems such as those considered in the paper, the importance of collaboration and learning the pertinent subject matter cannot be overemphasized. Agreed! Whether data are used wisely is in fact not only a statistical matter. A statistical application is good to the extent it is statistically sound and is also helpful for the subject matter field. Without basic language and understanding of the field the risk of irrelevance is high. In his Hunter lecture David provides three wise analyzes, as he also does in his many other applied papers. He has evidently an intimate knowledge of important areas of environmental science as well as the areas in biology, earth science and other fields where he has contributed, and he cooperates well with subject matter scientists.

The 2005 Neyman Lecture: Dynamic Indeterminism in Science [2008] Jerzy Neyman (1894 to 1981) founded the Statistics Department at Berkeley in 1938 after 4 successful but turbulent years in London following his early years in Poland. Neyman's life history and some of his contribu-

tions to applied statistics are reviewed, with emphasis on his use of dynamic stochastic modelling in his applied work in astronomy, fisheries science and weather modification. Neyman was concerned with phenomena developing in time and space. David briefly presnts stochastic differential equations (SDEs) as a background to Neyman's applied work, and as a common thread in his own applied work, not the least what he presents in the paper to expand on Neyman's work and to support the case for dynamic indeterminism in science. "Indeterministic" was for Neyman broadly synonymous with "stochastic" and "statistical". Chaos or other non-probabilistic indeterminism are not mentioned, perhaps for good reasons since Neyman was a practical man who sought empirical knowledge in the many fields of science where he worked. David is also a practical man, and he uses finite differences to obtain likelihood functions in his SDE models.

Two of the three examples of Neyman's applied statistics works were done together with Elizabeth Scott. Together with astronomers they developed the Neyman-Scott model for clustered point processes when studying the spatial distribution of galaxies. From graphical comparison of photographic images of the sky and images obtained by simulating their model, they found that more clustering was needed, and they developed a two-stage Neyman-Scott model. Weather modification was another area they investigated together. Through a randomized experiment of cloud seeding in Switzerland they discovered a "far-away" effect of increased rainfall far away from the seeding. The third study was done before Scott entered the Berkeley Department as a PhD. Here, Neyman developed and estimated an age-structured model for the population dynamics of Californian sardines that was subject to heavy exploitation. This study predates the seminal book: Beverton, R. J. H.; Holt, S. J. (1957) *On the Dynamics of Exploited Fish Populations.*

David follows up on Neyman's sardine study with his own study of sheep blowflies by way of a population matrix model. This is discussed above (Brillinger, Guckenheimer, Guttorp and Oster, 1980) . David expands on the weather modification study of Neyman and Scott by using a model for transforming a point process model of seed particles above Ticinino in Switzerland to a point process of rain drops in clouds blown to Zürich, and estimates the expected delay time from seeding to increased rainfall in Zürich. David's third example is a study of spatial motion of elks fitted with GPS collars in the enclosed Starkey Experimental Forest. An SDE model was employed and the velocity field of elk motion estimated. To what extend does recreational use of the area, i.e. driving ATVs, affect elk behavior? This was studied in

an experiment were a driven ATV was tracked and its trajectory introduced as an explanatory variable in the SDE model for the motion of an elk during the experimental period. The elk was significantly affected by the ATV, and the effect is graphically quantified. In his final example David studies the foraging behavior of Hawaiian monk seals fitted with GPS tags, by way of an SDE model cast in potential function form. This and the previous example is akin to Brillinger, Preisler and Ager (2001) discussed above. The potential function was modelled as a linear combination of basic functions, and was estimated by least squares. Synthetic plots, i.e. simulated tracks, from the fitted motion were found not unlike observed tracks where unreasonable satellite recorded position are cleaned up.

Sympathy and admiration for Neyman shines through the paper, but in his modest way David does not say how he was inspired and influenced by Neyman. From similarities in their appetite for applications to substantive sciences and in their great contributions, also to what we used to call mathematical statistics, the influence must have been substantial - or they were similarly gifted and born under the same star.