
Data Mining to distinguish Wave vs. Thermal Climbs in Soaring Flight Data

Alfred Ultsch and Rene Heise

Databionics Research Group
Philipps-University Marburg
ultsch@informatik.uni-marburg.de

Abstract

The accurate prediction of atmospheric waves and turbulence is a challenge for meteorological forecasting. To obtain precise measurement data on Mountain Waves is costly and only sparsely available. This work presents the exploitation of large databases of flights made by glider pilots for such purposes. The application of techniques from machine learning to extract useful data from such flights is presented. The quality of several different classifiers to distinguish Thermal Climbs from Wave Climbs is measured using supervised data sets from flights in the European Alps. The performance of different Bayesian Classifiers operating on Gauss Mixture Models of marginal variables is reported. A classifier using the same methods and relying on expert domain knowledge is constructed for the Andes. This is a region which is of high interest to meteorological research where, however, Ground Truth is hardly available.

1. Introduction

Strong winds blowing perpendicular to mountain ranges produce high reaching atmospheric waves on the lee side of the mountain, just as water waves on the back side of a submerged stone. The turbulence associated with such Mountain Waves can be a serious threat to the crews and passengers of commercial airline flights. Early theories on the occurrence and details of lee waves and associated turbulence were based on hydrological models, see (Doernbrack et al 2006) for an

overview. In recent years it has become apparent that these models need to be refined in order to allow more precise predictions of the occurrence of these turbulences (Dummann 2008). Empirical data is required to verify/falsify the meteorological models. Obtaining such empirical data is, however, not easy. Research flights with special equipped airplanes to measure lee waves are extremely costly (Hacker, J. et.al. (2007)). Furthermore, such flights can only measure some points in time and space of a phenomenon that may cover thousands of kilometers for more several hours. To obtain sufficient data another, broader approach may be useful.

The rising side of the waves allows gliders to gain altitude or fly long distances when soaring. So a substantial number of sport glider flights are performed in lee wave conditions all over the world. Almost all contemporary gliders are equipped with a GPS- (Global Positioning System) based flight recording system (logger). These recordings are published by the pilots for the purpose of a decentralized competition and also used to document record flights. In this work we demonstrate how machine learning algorithms can be used to extract meaningful data for research out of logger files produced by sport glider pilots. In particular the problem of assessing the performance of classifiers for is addressed for such areas of the world, where it is important that lee waves are studied but no Ground Truth is available.

2.The distinction of Thermals Climbs from Waves Climbs

One of the most interesting places to observe and measure lee waves are the Andes of South-America. This mountain range, with an average height of about 4,000 m forms a quite narrow North-South ridge perpendicular to the prevailing westerly wind direction. World record glider flights are flown using the lee waves in the Andes and produce valuable data. The so called Mountain Wave Project (MWP) systematically collects and analyzes such data (<http://mwp.flightplanner.info>). In GPS logger files the position of the glider, i.e. Longitude, Latitude, Altitude and Time is recorded at time intervals from one to four seconds. The sources of lift for a glider may either be updrafts caused by thermals (Thermal Climbs) or the upwind side of a lee wave (Wave Climbs). The distinction between these two types of climbs is essential for the research in Mountain Waves. In about 100 flights of the MWP measuring campaigns in the Andes 2.500 Climbs could be identified. For these Climbs the distances flown in the climbs (Distance Covered) and the entrance altitude (Climb Foot

Height) were collected. Figure 1 shows the distribution of the distances, Figure 2 the climb foot heights. Both distributions are measured using the Pareto Density Estimation (PDE) as probability density estimation (Utsch 2003). The figures also show in dashed lines a suitable Gaussian Mixture Model (GMM) that was fitted to the data.

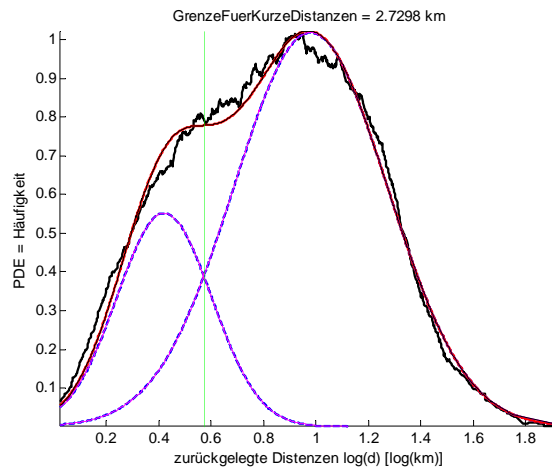


Figure 1: Distances covered in Climbs

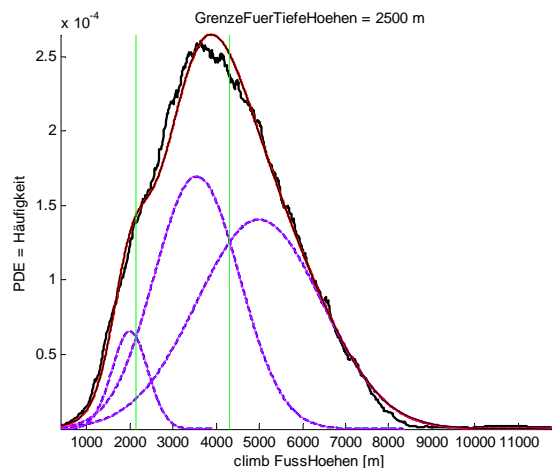


Figure 2: Climb entrance heights.

A plausible distinction between Thermal vs Wave climbs is that thermal climbs are more stationary. Thermals are locally (<2km) organized updrafts, whereas waves can stretch for more than 50 km. Thermals

occur in the lower, convective layers of the atmosphere. Mountain waves are typically organized in the mid to upper (>3km GND) layers. So it is reasonable to use a Bayes Classifier based on the marginal GMMs that assigns the climbs which start at lower altitudes and where only short distances are flown into the class of Thermal Climbs. Climbs which start at higher altitudes and where larger distances are covered are classified as Wave Climbs. In order to use the data for meteorological research an assessment on the quality of this knowledge driven unsupervised classifier for the Andes data should be made.

3. Assessing the Classification Quality

The data of the Andes are very valuable for the research on Mountain Waves since the orographic and meteorological conditions are very canonical: a narrow range of high ridges perpendicular to the wind. In this area of the world there are, however very few glider flights compared to, for example, the European Alps. For Europe it is also possible to obtain more meteorological data than in the sparsely populated Andes regions. In order to obtain verified classifications of climbs, Philip Ohrndorf, a glider pilot and student of Geography at the University of Marburg, searched the published flights of the years 2007-2009 for flights in the Alps that were partially flown in lee waves (Ohrndorf 2009). Ohrndorf identified 160 out of ca. 200.000 flights that had this property. Within each flight he classified the climbs into the categories Thermal Climbs vs. Waves Climbs. For this knowledge of the mountain shapes where the flight was made and the meteorological observations (wind, atmospheric conditions) were exploited. So ground truth data is available for these flights in the Alps. The data was 80/20 percent split into training and test data subsets. The measurements reported below are based on 100 fold cross validation.

For the variables “Horizontal Speed in a Climb”, “Distance Covered in Climbs”, “Climb Foot Height” and “Climb Top Height” a GMM on the variables as well as a multidimensional Gaussian Model (Multi-Gauss) was fitted. Preprocessing was square root (Sqrt) for heights and log for distances. The performance of the Bayesian Classifiers based on these models measured. Figure 3 shows the results.

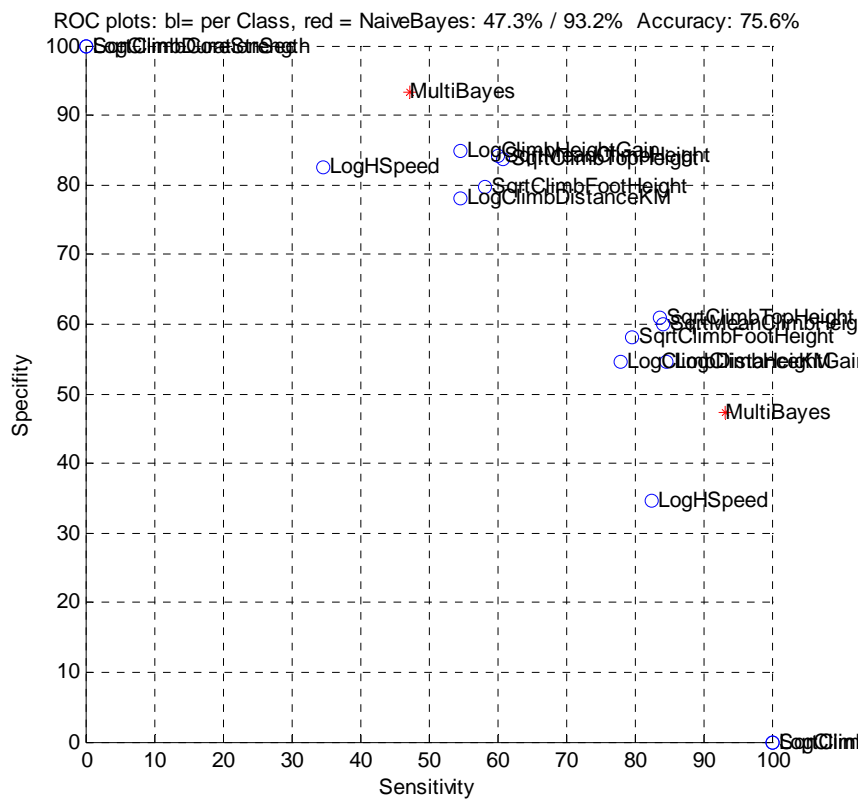


Figure 3: Sensitivity/Specificity of Bayes Classifiers

It can be seen that best classification results are obtained using the same variables as in the Andes classifier (Distance and Foot Altitude). Then the performance of four types of classifiers were measured on the data set: first a classifier constructed the same way as for the Andes data (see above), a Bayes Classifier using the independence assumption (Naïve Bayes Classifier), a tree based classifier using the CART algorithm and a Multilayer Perceptron (MLP/BP) with the back-propagation algorithm. The performance of the classifier are shown in Figure 4 (Ohrndorf 2009).

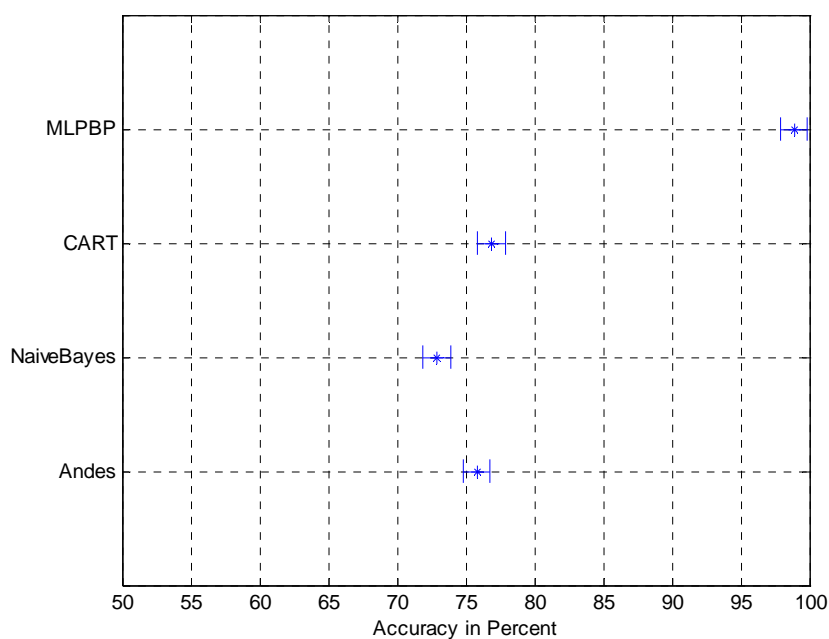


Figure 4: Performance of different classifiers on the Alps data

5. Discussion

The most interesting data for meteorological research on Mountain Waves and associated turbulence is the data from the measure campaigns in the South American Andes. For this data set, however hardly ground truth data is available to verify the performance of classifiers. For the same classification task data sets including a correct classification could be obtained in the European Alps. Bayes Classifiers on each marginal showed that the two variables “Climb Foot Height” and “Distance in Climb” are among the best variables for the construction of a classifier (Ohrndorf 2009). A combined classification of all variables (Multi-Bayes) does not show significant increase in performance (compare Figure 3). The performance measured for different classifiers (Figure 4) allows the conclusion that the construction principle used in the Andes Classifier yields comparable results to a Naïve Bayesian or a rule based Classifier. The variables

calculated by the CART classifier were almost the same as in the Andes Classifier.

Lee Waves in the Alps are much less organized as in the Andes for two primary reasons: first, the Alps are L-shaped with a N/S leg in France and an E/W leg from Switzerland to Austria. Cross sections of the Alps show several ridges of different orientation. This produces a more complex pattern of Mountain Waves than in the Andes. Therefore it is plausible to assume that the problem to distinguish between Thermal and Wave climbs is more difficult for Alps data than for the Andes data. So, with all necessary precautions, the conclusion could be drawn that the performance of the Andes classifier on the flight data of the Andes is on the same order of magnitudes as for the Alps data, i.e. in the 75% accuracy range.

The performance of the MLP/BP of 99% (see Figure 4) shows that better classifiers, as the Bayesian Classifiers can be constructed. Such a construction requires however that verified ground is available (Ohrndorf 2009). This type of classifiers, however, are subsymbolic i.e. the classification method is hidden and not determined by expert knowledge of the problem.

The Output (Model Output Statistic (MOS)) of state of the art numerical weather forecasts can be improved substantially using appropriate statistical interpolation and mixture of models of different scaling level (MOS-MIX) (Haalmann/Knuepfer 2009). However, a verification of these models using three dimensional measured data on lee waves is hardly performed (Schmidt et al 2008). The work presented here is among the first steps in the generation of high quality data in order to improve such meteorological forecasts. Advances in this area will ultimately prevent crew and passengers from injuries of in-flight turbulences and allow for a more precise prediction of lee waves that can be used for flights powered or subsidized by renewable solar energy (Heise, R. et.al. (2009).

6. Summary

The accurate prediction of Mountain Waves is a challenge for meteorological forecasting. To measure precise data on Mountain Waves to verify meteorological predictions is costly and only sparsely available. This work presents the first steps in the exploitation of large databases of recreational flights made by glider pilots all over the world. During these glider flights valuable data is recorded as a side effect of GPS logging of such flights. We present here the application of

techniques from machine learning to extract from such flights useful high quality data. The quality of several different classifiers to distinguish Thermal Climbs from Wave Climbs is measured using supervised data sets from flights in the European Alps. It is shown that Bayesian Classifiers working on Gauss Mixture Models of marginal variables perform on a satisfactory level. A classifier using the same methods and relying on expert domain knowledge is constructed for the Andes. This is a region where Ground Truth is hardly available, however measurements of this region is very important for the improvement of turbulence.

Acknowledgements

This work draws upon the expedition, flights and research done within the Mountain Wave Project (MWP) (<http://mwp.flightplanner.info>) as well as the data preparation and classification by Philip Ohrndorf.

References

1. Doernbrack, A., Heise, R. , Kuettner, J. (2006): Wellen und Rotoren, Promet, Nr 32, Vol 1-2, S. 18-24
2. Dummann, J. (2008): Report on Glider Pilot Activities to document Lee wave-Events in Northern Germany, Proc. XXIX OSTIV Congress, Lüsse-Berlin, Germany.
3. Forstner, B. (2000):, Untersuchung von Gebirgswellen durch Auswertung von Segelflug- und Radiosondendaten, Diplomarbeit, Universität Wien
4. Hacker, J. et.al. (2007): Measuring mountain waves and turbulence at up 12km altitude over the Andes in South America using an instrumented motorised glider, 14th National Australian Meteorological and Oceanographic Society (AMOS), Adelaide.
5. Haalman D., Knüpfner K., (2003): MOS-MIX: Integrated Statistical Interpretation of Multiple Numerical Models, European Conference on Applications of Meteorology,ECAM.
6. Heise, R. et.al. (2009): Weather Forecasting for Soaring Flight, World Meteorological Organisation (WMO), No. 1038, 76 p.
7. Ohrndorf, P. (2009) Die Identifikation von Leewellen mit Hilfe von Flugwegaufzeichnungen am Beispiel ausgewählter Segelflüge im Alpenraum, Examensarbeit, Universität Marburg.
8. Schmidt,T., Torre, A. de la, Wickert, J. (2008): Global gravity wave activity in the tropopause region from CHAMP radio occultation data, Geophysical Research Letters, VOL. 35.

9. Torre, A. de la, Alexander, P.(2005): Gravity waves above Andes detected from GPS radio occultation temperature profiles, *Geophysical Research Letters*, VOL. 32.
10. Ultsch, A. (2003): Pareto Density Estimation: A Density Estimation for Knowledge Discovery, in Baier D., Wernecke K.D. (Eds), *Innovations in Classification, Data Science, and Information Springer*, pp. 91-100.