DATING BAGPIPE DRONES BY INTERNAL DIMENSIONS

. . ..

Department of Statistics, Carnegie Mellon University

November 30, 2001

ABSTRACT. This paper documents an analysis of the internal dimensions of the drones of 23 highland bagpipes, collected by the bagpipe maker Roderick MacLellan. Several statistical models are presented which predict the year in which a set of drones were made, given the internal diameters of various chambers. This is apparently a new method of dating bagpipe drones.

0. INTRODUCTION

The owner of a bagpipe will generally wish to know who made the bagpipe and when. This may be simply to satisfy curiosity or may be for the more practical purposes of ordering replacement parts or setting a sale price for the bagpipe.

Determining the maker and age of a bagpipe can be difficult, as bagpipes often are not marked by their makers. An expert usually forms an opinion on the basis of external shape and decorative features of a bagpipe: the outside diameters and shape of the drones, the pattern of combing (grooves cut into the exterior surfaces of the drones and stocks) and the size, shape, material and condition of ornamental pieces (horn, ivory, silver, etc.).

Roderick MacLellan, a bagpipe maker in Lakewood, New Jersey, has recorded internal dimensions of the drones and stocks of twenty-three bagpipes which have passed through his workshop. These bagpipes were produced by a variety of makers over a wide range of time.

This paper summarizes an analysis of Mr. MacLellan's data. We show that the functional design of bagpipe drones has undergone a more-or-less steady change over the past 120 years. This change can be used to predict the year of production of a bagpipe from the internal physical dimensions of its drones.

Section 1 provides background information on the highland bagpipe. A full description of the data is in Section 2, and exploratory data analysis is described in Section 3. Methods of constructing and evaluating year-prediction models are presented in Section 4. Sections 5, 6 and 7 respectively contain year-prediction models based on a tenor drone, a bass drone and a complete bagpipe. Section 8 provides a summary of the best models, draws conclusions from them and suggests directions for future work.

The amount of bagpipe data available for this analysis is quite small; this causes some problems in constructing statistical models. The reader is asked to consider the models presented here as "proof of concept" rather than as finished products.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

1. Background

This section provides background information on the highland bagpipe and some heuristics which will be used in the subsequent analysis.

The Bagpipe.

The Great Highland Bagpipe (Figure 1.1) is a musical instrument consisting of three *drones*, single-reed pipes which produce a constant pitch, and a *chanter*, a double-reed pipe on which a melody is played. The drones and chanter fit into *stocks* attached to a bag. Constant air pressure in the bag causes the reeds to vibrate, producing sound from the chanter and drones.

Figure 1.1. A bagpipe by Roderick MacLellan [3].

A bagpipe has two identical tenor drones, each composed of two sections (*first* and *top*) and a single bass drone, composed of three sections (*first*, *middle* and *top*). Drone sections fit into one another as shown in Figure 1.2. The depth to which a drone section is pushed into the section above it affects the overall length of the drone and hence the fundamental frequency the drone produces.

A bagpipe drone is an acoustic filter. As Figure 1.2 shows, the internal diameter of the tenor drones changes four times and the internal diameter of the bass drone changes six times, dividing the drones into four or six *chambers*, respectively. The changes in diameter cause changes in impedance to sound waves; the changes in impedance cause partial reflection of the waves; the interaction of sound waves with the reflections causes constructive and destructive interference: thus the acoustic effect of the drone is determined by the lengths and internal diameters of its chambers.

The Data and Sources of Variation.

Variation by Design.

Each bagpipe maker will obviously have his or her own drone design, presumably that which in the maker's opinion gives the most pleasing tone. It would be very surprising if many chamber dimensions did not vary from maker to maker. Any maker's design would probably evolve over time.

It is generally held among pipers that the fundamental pitch of the bagpipe has risen over the last hundred years or so, from about 440Hz to around 470-480Hz. It is to be expected that some alterations in drone design have occurred due to this trend.

Variation by Tool Wear.

Bagpipes are made of extremely hard, dense wood, African Blackwood most commonly. Consequently, the tools used to make bagpipes undergo a continual process of dulling, re-sharpening and replacement, resulting in variation in tool shape and hence in finished bagpipes of any single design.

Variation by Age.

Wood changes dimensions over time, and two pieces which were identical when they were produced in 1900 may not be identical now.

Variation by Replacement.

Not all parts of a bagpipe drone may be original. Wood can crack over time, and replacement bagpipe parts (either re-made or scavenged from another incomplete bagpipe of the same make and approximate age) may not be identical to the originals. Occasionally the first section of the bass drone of an older bagpipe is replaced by a new piece with smaller internal diameter.[1] This is done to raise the pitch of the base drone slightly, accommodating the overall change in pitch mentioned above. Stocks are sometimes changed also.

2. The Data

Mr. MacLellan recorded the length and internal diameter of the chambers of the drones, more or less consistently according to his interest, on paper forms. From these forms, the author created the data file shown in Appendix 1. Mr. MacLellan did not make note of any replacement parts in his data, so it is assumed that all of these bagpipes were original.

[1]This will have implications later, when we decide whether to include the base drone first section internal diameter in prediction models.

FIGURE 1.2. Cartoon cross-sections of a bagpipe tenor drone (left) and bass drone (right). The 11 measurement variables used in this analysis, described in Section 2, are indicated.

The data file consists of twenty-nine variables. Four are used to identify the maker of the bagpipe and the year it was produced: the remaining twenty-five are measurements of the stocks, the two sections of the tenor drones and the three sections of the base drone. All physical measurements are recorded either in inches or in thousandths of an inch: it is clear from context which unit is used.

Not all measurements were recorded for all bagpipes. In the following list, each variable name is followed by the number of pipes for which that variable is recorded, 1 to 23. If a variable was recorded for 17 or more pipes it was considered in the analysis that follows: the names of these variables are written in italics below, and they are listed in Figure 1.2.

Identification. Name of the maker (*MAKER*, 23) and year of production (*YEAR*, 17). From comments on the forms, "confidence" variables for the maker (*MCON*, 23) and year of production (*YCON*, 23) were created, with 0 indicating that the maker or year is unknown, 1 indicating a low level of certainty and 2 indicating a high level of certainty.

Stock. Internal diameter (*STOCKID*, 17).

Tenor First Section. Internal diameter (*TFSID*, 23) and length (TFSLE, 7). For two pipes, the two tenor first sections have different diameters, recorded in the variable (TFSIDO, 2): the average of the two diameters is used in place of TFSID in these cases.

Tenor Top Section. Length (TTLE, 8), tuning chamber internal diameter (*TTTCID*, 23) and depth (TTTCDE, 4), second chamber internal diameter (*TT2CID*, 23), bell internal diameter (TTBEID, 1) and depth (TTBEDE, 1) and bush internal diameter (*TTBUID*, 21). For two pipes, the two tenor tuning chambers have different internal diameters, recorded in the variable (TTTCIDO, 2): the average of the two diameters is used in place of TTTCID in these cases.

Bass First Section. Internal diameter (*BFSID*, 21) and length (BFSLE, 8).

Bass Middle Section. Length (BMSLE, 8), tuning chamber internal diameter (*BMSTCID*, 22) and depth (BMSTCDE, 3) and second chamber internal diameter (*BMS2CID*, 18).

Bass Top Section. Length (BTLE, 9), tuning chamber internal diameter (*BTTCID*, 23) and depth (BTTCDE, 4), second chamber internal diameter (*BT2CID*, 23), bell internal diameter (BTBEID, 1) and depth (BTBEDE, 1) and bush internal diameter (*BTBUID*, 21).

3. Exploratory Data Analysis

Maker and Year of Production.

The twenty-three bagpipes in the data set were produced by thirteen different makers. The year of production is known, sometimes approximately and sometimes exactly, for all but six. The makers and the production years of their pipes are listed in Table 3.1. If the production year of a bagpipe is unknown, the years in which the maker was in business are given in Table 3.1, obtained from [1].

The Eleven Predictor Variables.

As mentioned in Section 2, eleven variables are observed frequently enough in the data that they are considered as predictors of the year of production. The distribution of these eleven variables are summarized in the side-by-side boxplots shown in Figure 3.1.

From Figure 3.1, it appears that the largest internal diameter is that of the stocks (STOCKID) and that the rest are grouped by function. Next in decreasing order of size are the three tuning chambers (TTTCID, BMTCID and BTTCID), followed by the two bushes (TTBUID and BTBUID). The three second chambers come next, with the two top second chambers (TT2CID and BT2CID) being similar

Maker	Years Represented	Years in Business
Lawrie	1914,1914,1945,?,?	1881 – 1980s
Henderson	1880,1924,1924,?,?	1880 – 1973
Hardie	1950,1957,1960	
Sinclair	1956	
Starck	1905	
Fletcher	1997	
Grainger and Campbell	1952	
Naill	1990	
J. Glen	1890	
Kintail	1981	
MacDougall	?	1792 – 1919
Gibson	1987	
Robert Reid	?	1932 – 1957

TABLE 3.1. The 13 makers and the corresponding production years of their pipes in the data set.

in size and the bass middle section second chamber (BMS2CID) clearly smaller. Smallest are the first section diameters (TFSID and BFSID).

Good Predictors of Year of Production.

Scatterplots and simple linear regressions were studied to determine whether any of the eleven variables mentioned above are related to the year in which a bagpipe was produced. Five appear to be: the internal diameters of the stocks (STOCKID), the tenor first sections (TFSID), both chambers of the bass middle section (BMSTCID and BMS2CID) and the bass top section tuning chamber (BTTCID). Details of the simple linear regressions described in this section may be found in Appendix 2.[2]

Categorization by Age.

A relationship between some of these variables and YEAR only becomes apparent when the bagpipes are categorized by age. Three age categories are imposed: two *antique* bagpipes, made before 1900; eleven *old* bagpipes, made between 1900 and 1980; four *new* bagpipes, made after 1980. The Robert Reid bagpipe (**23**)[3] is certainly old in this sense, as are probably all of the other bagpipes of unknown age except the MacDougall (**21**), which is probably antique.

Stock Internal Diameter — STOCKID.

A scatterplot of YEAR against STOCKID is shown in Figure 3.2, along with the fitted least-squares line. The slope and intercept of the line are -0.79 and 2559.97: the internal diameter of the stocks appears to be decreasing with time, at about 0.79 thousandths of an inch per year. The slope is significantly different from zero at the 5% level.

Tenor First Section Internal Diameter — TFSID.

A scatterplot of YEAR against TFSID is shown in Figure 3.3. The point on the extreme left of the plot, separated from the rest of the data, is the Sinclair bagpipe (**3**). This bagpipe has small internal diameters generally, and a dramatically small tenor first section internal diameter.

[2] In addition to regressing YEAR on the internal diameter variables, YEAR was also regressed on the square of each: essentially, cross-sectional area was investigated as a predictor of YEAR. No significant linear relationship between YEAR and any cross-sectional area was found.

[3] Bold numbers in parentheses are the index of a bagpipe in the data file of Appendix 1.

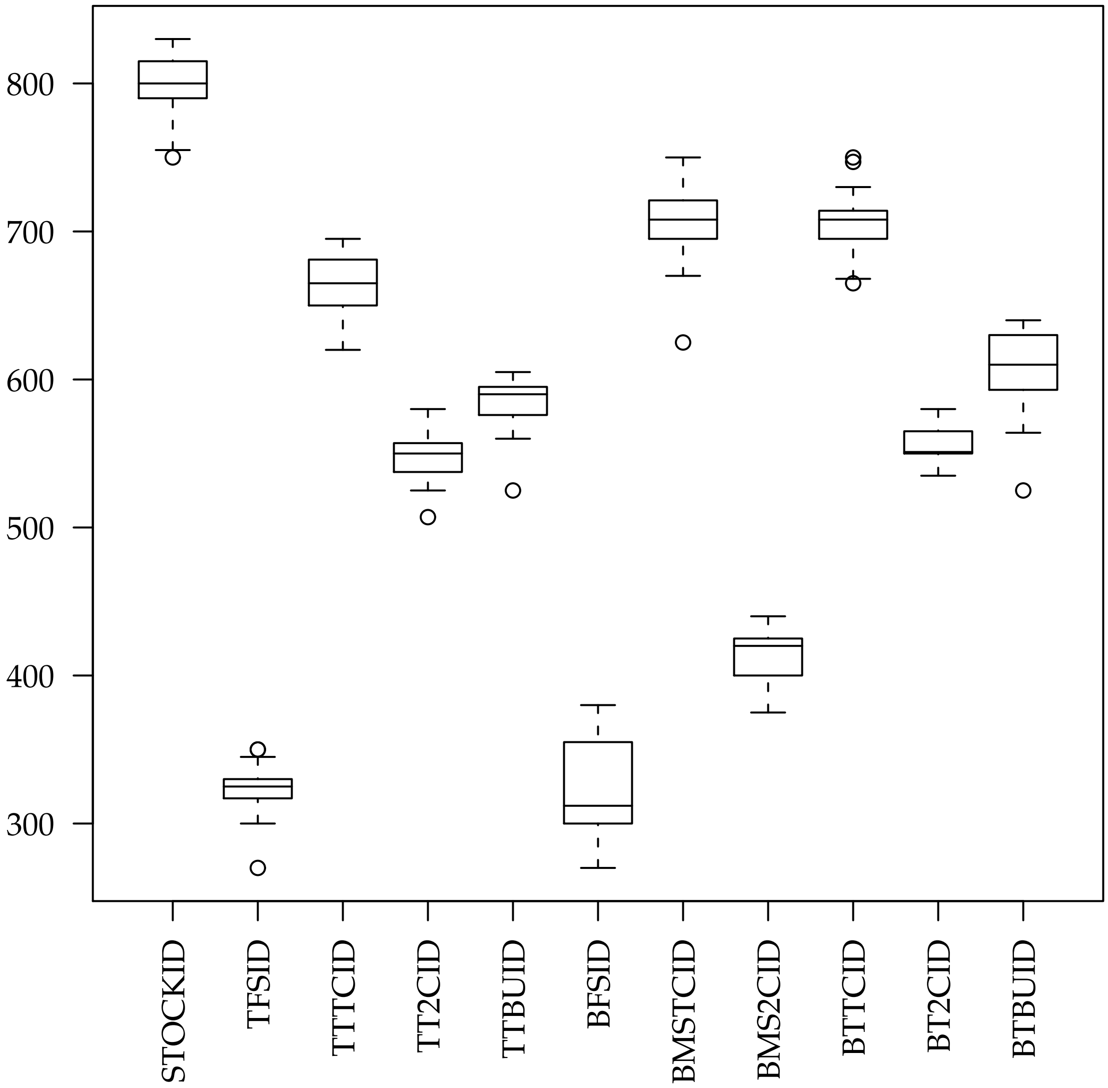

FIGURE 3.1. Side-by-side boxplots of the eleven predictor variables for YEAR. The vertical axis is internal diameter, in thousandths of an inch. The two very low bush diameters (TTBUID and BTBUID) come from the Grainger and Campbell pipe (**11**). No other pattern is evident among the outliers in these plots.

Two fitted least-squares lines are shown in Figure 3.3. The dashed one was fitted using all of the data and the solid one was fitted while excluding the Sinclair bagpipe (**3**). The slope and intercept of the solid line are -1.76 and 2520.28: the internal diameter of the tenor first section appears to be decreasing with time, at about 1.76 thousandths of an inch per year. The slope is significantly different from zero at the 1% level.

The four new bagpipes are clustered slightly above the least-squares line and the two antique bagpipes are clustered slightly below it. Removing these and regressing YEAR on TFSID for only the old bagpipes, with or without the Sinclair, also produces good linear fits of the data.

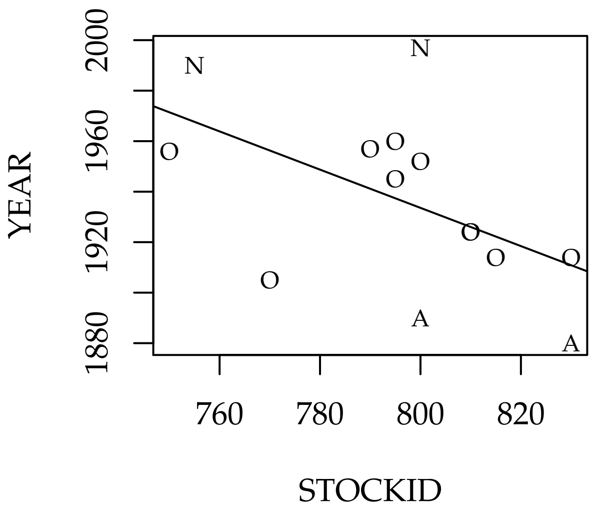

FIGURE 3.2. YEAR plotted against STOCKID. The least-squares line is shown. New bagpipes are represented by "N", old bagpipes by "O" and antique bagpipes by "A".

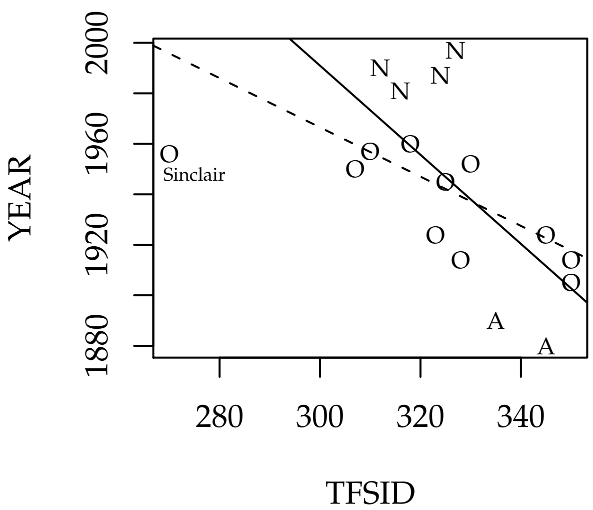

FIGURE 3.3. YEAR plotted against TFSID. Two least-squares lines are shown: the dashed line was fitted using all of the data and the solid one was fitted while excluding the Sinclair bagpipe.

Bass Middle Section Tuning Chamber Internal Diameter — BMSTCID.

A scatterplot of YEAR against BMSTCID is shown in Figure 3.4. Three of the four new bagpipes and the two antique bagpipes have very similar bass middle section tuning chamber internal diameters, approximately 700 thousandths of an inch. They are at the top and bottom of the plot, respectively. The points corresponding to old bagpipes approximately lie on a line.

A fitted least-squares line is shown in Figure 3.4: this was fitted using only the old bagpipes. The slope and intercept of the line are -0.68 and 2417.95: for old bagpipes (produced between 1900 and 1980), the internal diameter of the bass drone middle section tuning chamber appears to be decreasing with time, at about 0.68 thousandths of an inch per year. The slope is significantly different from zero at the 1% level.

One of the four new bagpipes, the Gibson (**22**) lies quite close to the fitted line, at the extreme left of the plot. It has followed the apparent trend of the old bagpipes, while the other new pipes have not.

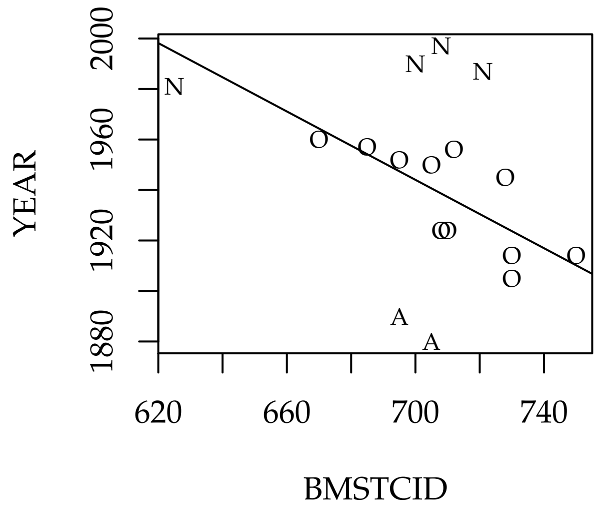

FIGURE 3.4. YEAR plotted against BMSTCID. The least-squares line computed using only the old bagpipes is shown. One new bagpipe, the Gibson (**22**) lies close to this line; the other new and antique pipes lie far away from it.

Bass Middle Section Second Chamber Internal Diameter — BMS2CID.

A scatterplot of YEAR against BMS2CID is shown in Figure 3.5. As with BMSTCID, three of the four new bagpipes differ from a linear trend evident in the old bagpipes: the antique pipes appear to follow the trend.

A fitted least-squares line is shown in Figure 3.5: this was fitted using only the old and antique bagpipes. The slope and intercept of the line are -1.20 and 2419.55: for old and antique bagpipes (produced before 1980), the internal diameter of the bass drone middle section second chamber appears to be decreasing with time, at about 1.20 thousandths of an inch per year. The slope is significantly different from zero at the 1% level.

The Gibson bagpipe (**22**), which is new but followed the older trend for BMSTCID, does not appear in this plot because its BMS2CID is unknown.

Bass Top Tuning Chamber Internal Diameter — BTTCID.

A scatterplot of YEAR against BTTCID is shown in Figure 3.6. The plot has much in common with that of BMSTCID (Figure 3.4), in that the four new bagpipes and the two antique bagpipes form two groups standing apart from a linear trend among the old bagpipes.

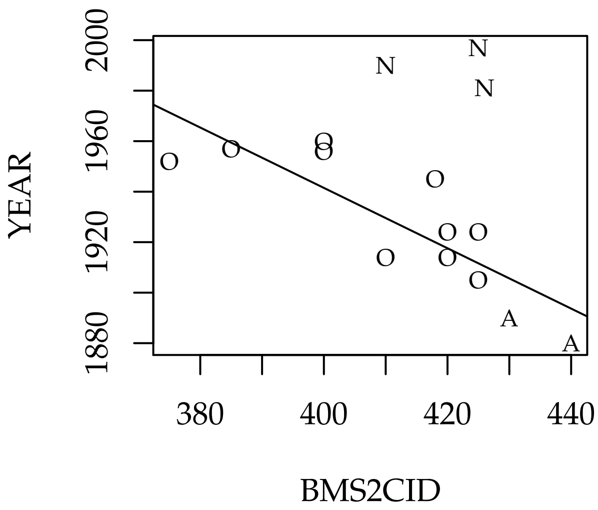

FIGURE 3.5. YEAR plotted against BMS2CID. The least-squares line computed using only the old and antique bagpipes is shown.

A fitted least-squares line is shown in Figure 3.6: this was fitted using only the old bagpipes. The slope and intercept of the line are -0.79 and 2493.48: for old bagpipes (produced between 1900 and 1980), the internal diameter of the bass drone top section tuning chamber appears to be decreasing with time, at about 0.79 thousandths of an inch per year (the same rate at which stock internal diameter is decreasing, to two decimals of accuracy). The slope is significantly different from zero at the 1% level.

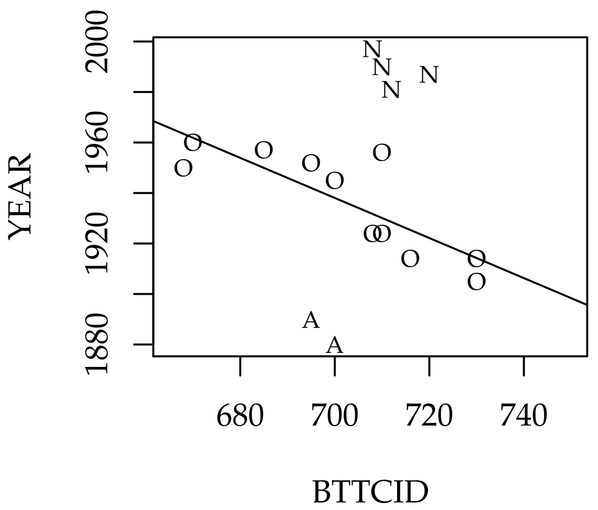

FIGURE 3.6. YEAR plotted against BTTCID. The least-squares line computed using only the old bagpipes is shown.

4. Overview of Models to Predict Year of Production

Variables.

Main Effects.

The main effects that will be considered are the eleven continuous variables discussed in Section 3, plus two dummy variables for the approximate age of a bagpipe, *DANT*[4] (1 if the pipe was made before 1900, 0 otherwise) and *DNEW* (1 if the pipe was made after 1980, 0 otherwise). All models will have an intercept.

The intended application of a model determines which main effects it may include. A model which is designed to date a complete bagpipe with all original parts might include the variable STOCKID, for example, but this model would be useless in dating a bagpipe with replacement drone stocks.

In this paper we will focus on three classes of models, respectively designed for dating only tenor drones (Section 5), only a bass drone (Section 6), and a complete bagpipe (Section 7). Within each class, it is desirable to consider models where STOCKID, approximate age (DANT and DNEW) or BFSID are unknown in addition to models in which all are known. In the interest of brevity we will consider only models in which all of these variables are known.

Interactions.

All pairwise interactions of main effects will be considered for inclusion in models, except those involving DANT and DNEW. For a complete-bagpipe model, then, there are $13 + \binom{11}{2} = 68$ variables to consider[5]. We will ignore the Principle of Marginality, which states that a model which contains an interaction should contain the main effects of that interaction.

Evaluating Models.

How are models to be compared? Since our intended application is prediction, it makes sense to choose a model which minimizes the mean squared prediction error (MSPE),

$$\text{MSPE} \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_{(i)})^2 \;,$$

where n is the size of the data set and $\widehat{y}_{(i)}$ is the prediction of y_i by the model fit on all but the i^{th} data point. This proportional to the "leave one out" cross-validation score.

Examples in Appendix 3 show that models which have near-minimum MSPE over some class tend to have near-minimum Akaike Information Criterion (AIC) over that class, and vice versa. The AIC is defined by

$$\text{AIC} \;=\; -n\,\ln\left(\frac{n}{\text{RSS}}\right) + 2k \;,$$

where RSS is the residual sum of squares resulting from the model and k is the number of variables in the model (so low values are good). The AIC is used for convenience, because the R software package contains a stepwise model selection routine which evaluates models by AIC.

In addition to the MSPE and AIC, the adjusted R^2 measure of fit will be used to guard against models which under-fit the data.

[4] In Section 5, 6 and 7 we will predict the year of production for some bagpipes for which YEAR is unknown using models which may contain DANT and DNEW. For such purposes, it is assumed that the MacDougall pipe (**21**) of unknown year is antique and that all other pipes of unknown year are old.

[5] Interactions involving DANT and DNEW are excluded for ease of model interpretation and to reduce the total number of variables. Interactions of three or more main effects are excluded only to reduce the number of variables. Allowing all interactions, there would be $2^{11} - 1 + 2 = 2049$ variables to consider when constructing a complete-bagpipe model.

Constructing Models.

Three general methods of constructing linear models are used: exhaustive search of low-rank[6] models, stepwise model selection (a subexhaustive search of the set of all possible models), and stepwise selection applied to models based on principal components.

Exhaustive Search of Low Rank Models.

Since we have more variables (up to 68) than data points (no more than 23), many models will over-fit the data. One way to avoid over-fitting is to impose a restriction ρ on rank and to compute the MSPE for every model with rank less than or equal to ρ. This method is computationally expensive (all the more so because it was done in R rather than in C), and our choice of ρ will be partly determined by computational feasibility. Appendix 3 contains the technical details of the exhaustive model fitting performed for Sections 5, 6 and 7.

Stepwise Model Selection.

The low-rank models with near-minimum AIC or MSPE will be used as starting points for stepwise model selection, as implemented by Venebles and Ripley's `StepAIC` routine in the "MASS" R library. The `stepAIC` routine halts when it has found a model with the property that adding or subtracting any one variable does not decrease the AIC. The algorithm also halts when it has found a model with AIC $= -\infty$, which is a sure sign that the model is over-fit! We shall have more to say about this in Section 5.

Principal Component Analysis.

The eleven predictor variables are highly collinear, just as one would expect the internal diameters of nested tubes to be. It is natural in this situation to attempt to reduce the number of variables by computing principal components. Appendix 4 contains the technical details of fitting models based on principal components for Sections 5, 6 and 7.

Comparing Models.

There are six bagpipes (**7**, **8**, **14**, **16**, **21**, **23**) for which YEAR is unknown. Different models will be applied to predict YEAR for these bagpipes when their known variables coincide with the variables used by the models. A further check on models is comparison of predictions of tenor-only models with bass-only models.

Heuristics.

The EDA shows that five main effects (STOCKID, TFSID, BMSTCID, BMS2CID and BTTCID) are significantly decreasing with time, so it is to be expected that some of these will appear in any good model. These five variables are highly correlated, so it would be surprising if all five appeared in a good model. It was also seen in the EDA that the relationship of some of these variables with YEAR is much more significant when the approximate age of the bagpipe is known, so it is to be expected that DANT and DNEW will appear in many good models. All other things being equal, we would like to avoid models which contain STOCKID, BFSID, DANT and DNEW, because these seem the likeliest variables to be missing for some bagpipes.

[6] In this paper, the *rank* of a linear model is the number of variables it contains. The set of all subsets of the variables, partially ordered by inclusion, is a ranked lattice where rank is equal to the size of the subset. By identifying a model with its variables, the set of all possible models is also a ranked lattice. Barring perfect collinearity in the data and including in intercept in all models, the rank of a model as defined here is one less than the rank of its design matrix.

5. Tenor Drone Models

The YEAR, STOCKID and all of the tenor drone variables are known for 13 bagpipes (**1**, **2**, **3**, **4**, **5**, **6**, **9**, **10**, **11**, **12**, **13**, **15**, **18**).

Exhaustive Search.

The best model found by exhaustive search of the 21776 models of rank less than or equal to 6 is

```
YEAR ~ STOCKID * TTBUID + TFSID + dant + dnew
```

This model had both the minimum MSPE (35.80) and AIC (50.44). In R format, the estimated coefficients of this model are

```
fit_lm(YEAR~STOCKID*TTBUID+TFSID+dant+dnew)
summary(fit)
                  Estimate Std. Error t value Pr(>abs(t))
(Intercept)    -1.926e+04  3.104e+03  -6.203 0.000809 ***
STOCKID         2.690e+01  3.874e+00   6.942 0.000443 ***
TTBUID          3.716e+01  5.435e+00   6.837 0.000481 ***
TFSID          -6.472e-01  1.055e-01  -6.133 0.000860 ***
dant           -2.388e+01  7.536e+00  -3.170 0.019329 *
dnew            6.133e+01  5.021e+00  12.214 1.83e-05 ***
STOCKID:TTBUID -4.667e-02  6.783e-03  -6.880 0.000465 ***
Residual standard error: 5.978 on 6 degrees of freedom
Multiple R-Squared: 0.9842, Adjusted R-squared: 0.9684
F-statistic: 62.21 on 6 and 6 DF,  p-value: 3.867e-05
```

The high adjusted R^2 suggests that this model is not under-fitting the data, the coefficients are all significant and the residual plots for this model (not shown) look quite good. Except for TTBUID, this model is composed of variables identified as significantly related to YEAR in the EDA.

For two bagpipes with unknown YEAR (**14** and **23**), all of the model variables are known. Using this model to predict YEAR for these two pipes results in 95% prediction intervals:[7]

$$\text{Bagpipe } \mathbf{14} \rightarrow (1923.703, 1940.318, 1956.932)$$

and

$$\text{Bagpipe } \mathbf{23} \rightarrow (1917.696, 1934.041, 1950.386)\ .$$

These predictions are consistent with the years in which Henderson (**14**) and Reid (**23**) were in business, shown in Table 3.1.

Stepwise Search.

A stepwise search algorithm was started at each of the ten best models constructed by the exhaustive method described above, ranked by MSPE. These models are listed in Appendix 3.

Finite AIC.

Of the ten initial models, three produce search paths which converge to models with finite AIC. One of these initial models is a local minimum in AIC; a stepwise search begun there stays there. The other two search paths converge to the rank 8 model

```
YEAR ~ (STOCKID*TFSID*TTBUID - STOCKID:TFSID:TTBUID) + dnew + dant
```

[7] The notation here means that the predicted year of production of bagpipe (**14**) is 1940, and that we are 95% confident that is was produced between 1923 and 1957.

This model has MSPE = 19.39 and AIC = 45.95. In R format, the estimated coefficients of this model are

```
fit_lm(YEAR~(STOCKID*TFSID*TTBUID-STOCKID:TFSID:TTBUID)+dnew+dant,x=T)
summary(fit)
                 Estimate Std. Error t value Pr(>abs(t))
(Intercept)    -1.715e+04  3.080e+03  -5.568 0.005098 **
STOCKID         3.310e+01  4.759e+00   6.955 0.002246 **
TFSID          -7.617e+00  4.025e+00  -1.892 0.131382
TTBUID          2.557e+01  9.095e+00   2.811 0.048266 *
dnew            5.716e+01  5.297e+00  10.791 0.000418 ***
dant           -1.836e+01  8.142e+00  -2.255 0.087165 .
STOCKID:TFSID  -1.813e-02  1.212e-02  -1.496 0.208981
STOCKID:TTBUID -4.725e-02  6.742e-03  -7.009 0.002181 **
TFSID:TTBUID    3.662e-02  2.039e-02   1.796 0.146898
Residual standard error: 5.282 on 4 degrees of freedom
Multiple R-Squared: 0.9918, Adjusted R-squared: 0.9753
F-statistic: 60.21 on 8 and 4 DF,  p-value: 0.0006671
```

The residual plots for this model look quite good, though the Sinclair bagpipe (**3**) looks as though it has high influence and the residual of (**18**) is about twice as large as the others. This model is composed of the same main effects as the best model found by exhaustive search.

For two bagpipes with unknown YEAR (**14** and **23**), all of the model variables are known. Using this model to predict YEAR for these two pipes results in 95% prediction intervals:

$$\text{Bagpipe } \mathbf{14} \rightarrow (1923.779, 1940.438, 1957.098)$$

and

$$\text{Bagpipe } \mathbf{23} \rightarrow (1916.217, 1933.737, 1951.256)\ .$$

These predictions are almost identical to those produced by the best model found by exhaustion.

Infinite AIC.

Of the ten initial models, seven produced search paths which converged to models with AIC $= -\infty$, that is, models with as many variables as data points. This is an artifact of the small data size: clearly, these models with are overfit.

It is interesting to examine the search path from a starting model to an overfit model. In the notation of Appendix 3, the stepwise search executed by

```
fit1_lm(y~X[,4]+X[,5]+X[,6]+X[,8]+X[,10]+X[,11])
fit2_stepAIC(fit1,scope=list(upper= ~X[,1]+X[,2]+...+X[,17],lower= ~1))
```

produces the (very truncated) output

```
AIC= 52.56
y~X[,4]+X[,5]+X[,6]+X[,8]+X[,10]+X[,11]+X[,17]
Step:AIC=47.74
y~X[,4]+X[,5]+X[,6]+X[,8]+X[,10]+X[,11]+X[,17]+X[,14]
Step:AIC=28.45
y~X[,4]+X[,5]+X[,6]+X[,8]+X[,10]+X[,11]+X[,17]+X[,14]+X[,2]
Step:AIC=15.41
y~X[,4]+X[,5]+X[,6]+X[,8]+X[,10]+X[,11]+X[,17]+X[,14]+X[,2]+X[,9]
Step:AIC=-51.48
y~X[,4]+X[,5]+X[,6]+X[,8]+X[,10]+X[,11]+X[,17]+X[,14]+X[,2]+X[,9]+X[,16]
Step:AIC=-Inf
y~X[,4]+X[,5]+X[,6]+X[,8]+X[,10]+X[,11]+X[,17]+X[,14]+X[,2]+X[,9]+X[,16]+X[,1]
```

Translating the notation of Appendix 3, the starting model and next-to-last model in the search path are respectively

```
YEAR ~ TT2CID + TTBUID + dnew + STOCKID:TFSID + STOCKID:TT2CID + STOCKID:TTBUID
```

and

```
YEAR ~ TT2CID + TTBUID*TFSID + dnew + STOCKID:TT2CID + STOCKID:TFSID +
          STOCKID:TTTCID + STOCKID:TTBUID + TTTCID:TTBUID + TT2CID:TTBUID
```

The next-to-last model has MSPE = 19.39 and AIC = -51.48. In R format, the estimated coefficients of this model are

```
fit_lm(YEAR~TT2CID+TTBUID*TFSID+dnew+STOCKID:TT2CID+STOCKID:TFSID+STOCKID:TTTCID+
                 STOCKID:TTBUID+TTTCID:TTBUID+TT2CID:TTBUID,x=T)
summary(fit)
                 Estimate Std. Error t value Pr(>abs(t))
(Intercept)    -4.867e+04  9.380e+02  -51.89  0.01227 *
TT2CID          5.360e+01  1.558e+00   34.40  0.01850 *
TTBUID          1.345e+02  2.009e+00   66.94  0.00951 **
TFSID          -1.873e+01  4.561e-01  -41.07  0.01550 *
dnew            3.595e+01  6.657e-01   54.00  0.01179 *
TT2CID:TTBUID  -2.163e-01  3.779e-03  -57.23  0.01112 *
TT2CID:STOCKID  9.188e-02  8.990e-04  102.20  0.00623 **
TTBUID:TFSID    7.587e-02  1.312e-03   57.84  0.01100 *
TTBUID:STOCKID -5.971e-02  5.578e-04 -107.06  0.00595 **
TTBUID:TTTCID   1.132e-02  8.018e-04   14.12  0.04502 *
TFSID:STOCKID  -3.261e-02  5.054e-04  -64.52  0.00987 **
STOCKID:TTTCID -8.790e-03  5.962e-04  -14.74  0.04311 *
Residual standard error: 0.1978 on 1 degrees of freedom
Multiple R-Squared:     1, Adjusted R-squared:     1
F-statistic: 3.15e+04 on 11 and 1 DF,  p-value: 0.004395
```

The residuals do not appear to be normally distributed and an adjusted R^2 value of 1.0 is suspiciously high.

Using this model to predict YEAR for bagpipes (**14** and **23**) results in 95% prediction intervals:

$$\text{Bagpipe } \mathbf{14} \rightarrow (1942.471, 1949.894, 1957.317)$$

and

$$\text{Bagpipe } \mathbf{23} \rightarrow (1920.551, 1928.290, 1936.029)\ .$$

This model predicts that the Reid (**23**) bagpipe was made four years before Reid started making bagpipes (see Table 3.1)! We should be wary of using this model for prediction, as it is probably overfit.

Which model in the sequence above is best for prediction? There is a bias-variance trade-off problem to be studied here: as its relevance decreases as the data size increases, we choose not to address it, and search paths which lead to over-fit models will be ignored.

Principal Components.

The best model found using principal components is

```
y ~ p[,1] + p[,4] + p[,7]
```

where p[,1], p[,4] and p[,7] are the first, fourth and seventh principal components (the loadings of these components have no clear interpretation: see Appendix 4). This model has MSPE = 241.25 and AIC = 76.01. In R format, the estimated coefficients of this model are

```
fit2_lm(y~p[,1]+p[,4]+p[,7],x=T)
summary(fit2)
Coefficients:
            Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1939.846      4.559 425.476  < 2e-16 ***
   p[, 1]      7.039      1.548   4.547  0.00139 **
   p[, 4]    -12.611      3.414  -3.694  0.00497 **
   p[, 7]     22.453      8.591   2.614  0.02810 *
Residual standard error: 16.44 on 9 degrees of freedom
Multiple R-Squared: 0.8205, Adjusted R-squared: 0.7607
F-statistic: 13.72 on 3 and 9 DF,  p-value: 0.00105
```

The quality of fit of this model is disappointing, compared with the best models found by the exhaustive and stepwise methods.

6. Bass Drone Models

The YEAR, STOCKID and all of the bass drone variables are known for 14 bagpipes (**1**, **2**, **3**, **4**, **5**, **6**, **9**, **10**, **11**, **12**, **13**, **15**, **18**, **19**).

Exhaustive Search.

The best model found by exhaustive search of the 31929 models of rank less than or equal to 4 is

```
YEAR ~ dant + dnew + BMSTCID:BTBUID + BMS2CID:BTTCID
```

This model has both the minimum MSPE (47.25) and AIC (61.15). In R format, the estimated coefficients of this model are

```
fit_lm(YEAR~dant+dnew+BMS2CID:BTBUID+BMS2CID:BTTCID)
summary(fit)
                 Estimate Std. Error t value Pr(>abs(t))
(Intercept)     2.265e+03  4.095e+01  55.307 9.06e-14 ***
dant           -3.632e+01  6.161e+00  -5.895 0.000152 ***
dnew            7.146e+01  5.518e+00  12.951 1.42e-07 ***
BMS2CID:BTBUID  5.380e-04  1.825e-04   2.948 0.014578 *
BMS2CID:BTTCID -1.605e-03  2.388e-04  -6.722 5.22e-05 ***
Residual standard error: 7.402 on 10 degrees of freedom
Multiple R-Squared: 0.9691, Adjusted R-squared: 0.9567
F-statistic: 78.38 on 4 and 10 DF,  p-value: 1.65e-07
```

The high adjusted R^2 suggests that this model is not under-fitting the data, and the coefficients are all significant. The residual plots show that the residual for the Sinclair pipe (**3**) is unusually large. The reason for this is that the Sinclair pipe has fairly low values of each of BMS2CID, BTTCID and BTBUID, resulting in very low values for the interactions. Re-fitting this model without the Sinclair pipe in the data, the estimated coefficients of this model are

```
dant_dant[-3]; dnew_dnew[-3]
fit_lm(YEAR~dant+dnew+BMS2CID:BTBUID+BMS2CID:BTTCID,data=data[-3,])
summary(fit)
                  Estimate Std. Error t value Pr(>abs(t))
(Intercept)     2.257e+03  2.595e+01  86.954 1.78e-14 ***
dant           -3.480e+01  3.910e+00  -8.902 9.34e-06 ***
dnew            7.337e+01  3.518e+00  20.855 6.28e-09 ***
BMS2CID:BTBUID  5.549e-04  1.154e-04   4.810  0.00096 ***
BMS2CID:BTTCID -1.598e-03  1.509e-04 -10.590 2.22e-06 ***
Residual standard error: 4.676 on 9 degrees of freedom
Multiple R-Squared: 0.9887, Adjusted R-squared: 0.9837
F-statistic:   197 on 4 and 9 DF,  p-value: 9.418e-09
```

The residual plots for this model look much better than for the previous one. Except for BTBUID,[8] this model is composed of variables identified as significantly related to YEAR in the EDA.

For two bagpipes with unknown YEAR (**14** and **21**), all of the model variables are known. Using this model to predict YEAR for these two pipes results in 95% prediction intervals:

$$\text{Bagpipe } \mathbf{14} \rightarrow (1919.329, 1931.186, 1943.043)$$

and

$$\text{Bagpipe } \mathbf{21} \rightarrow (1872.413, 1886.764, 1901.115)\ .$$

These predictions are consistent with the years in which Henderon (**14**) and MacDougall (**21**) were in business, shown in Table 3.1. We expect to do badly using this model to predict YEAR for the Sinclair bagpipe. Doing so results in the 95% prediction interval:

$$\text{Bagpipe } \mathbf{3} \rightarrow (1924.992, 1936.180, 1947.368)\ .$$

This bagpipe is known to have been produced in 1956, so this prediction is off by 20 years!

Stepwise Search.

A stepwise search algorithm was started at each of the ten best models constructed by the exhaustive method described above, ranked by MSPE. These models are listed in Appendix 3.

All of the ten initial models produce search paths which converge to models with finite AIC. Four of these search paths converge to the model with lowest AIC, the rank 9 model

```
YEAR ~ STOCKID*BTTCID + dnew + dant + STOCKID:BMS2CID + BMSTCID:BTBUID +
            BMS2CID:BTTCID + BFSID:BMSTCID
```

This model has MSPE = 30.51 and AIC = 48.22. In R format, the estimated coefficients of this model

[8] It is interesting that for both the tenor and bass drones, the bush internal diameter (TTBUID and BTBUID) features significantly in the best models, in interactions. Neither TTBUID nor BTBUID appeared to be significantly related to YEAR in the EDA, where they were considered only as main effects.

are

```
fit_lm(YEAR~STOCKID*BTTCID+dnew+dant+STOCKID:BMS2CID+BMSTCID:BTBUID+
                 BMS2CID:BTTCID+BFSID:BMSTCID,x=T)
summary(fit)
                 Estimate Std. Error t value Pr(>abs(t))
(Intercept)     -1.058e+04  6.884e+03  -1.537 0.199197
STOCKID          1.326e+01  8.615e+00   1.539 0.198607
BTTCID           2.146e+01  1.013e+01   2.119 0.101447
dnew             7.443e+01  6.839e+00  10.883 0.000405 ***
dant            -3.369e+01  1.020e+01  -3.302 0.029870 *
STOCKID:BTTCID  -2.289e-02  1.199e-02  -1.909 0.128887
STOCKID:BMS2CID  7.106e-03  4.653e-03   1.527 0.201418
BTTCID:BMS2CID  -9.861e-03  5.192e-03  -1.899 0.130326
BMSTCID:BTBUID   5.472e-04  1.797e-04   3.045 0.038215 *
BMSTCID:BFSID   -1.022e-04  8.090e-05  -1.263 0.275162
Residual standard error: 5.125 on 4 degrees of freedom
Multiple R-Squared: 0.9934, Adjusted R-squared: 0.9785
F-statistic: 66.65 on 9 and 4 DF,  p-value: 0.0005347
```

The residual plots for this model look quite good. This model is contains main effects which are not in the best model found by exhaustive search.

Only for bagpipe (**14**) are all of the model variables are known and YEAR unknown. Using this model to predict YEAR for this pipe results in a 95% prediction interval:

$$\text{Bagpipe } \mathbf{14} \rightarrow (1914.038, 1936.826, 1959.613)\ .$$

This prediction is closer to those produced by the tenor drone models than that produced by the best bass model found by exhaustion.

Principal Components.

The best model found using principal components is

```
y ~ p[,1] + p[,3] + p[,4] + p[,5] + p[,6] + p[,8]
```

where `p[,i]` is the i^{th} principal component (the loadings of these components have no clear interpretation: see Appendix 4). This model has MSPE = 59.1 and AIC = 60.44. In R format, the estimated coefficients of this model are

```
p1_p[,1]; p3_p[,3]; p4_p[,4]; p5_p[,5]; p6_p[,6]; p8_p[,8];
fit2_lm(y~p1+p3+p4+p5+p6+p8,x=T)
Coefficients:
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1936.2857     1.9847 975.607  < 2e-16 ***
p1             4.8797     0.5228   9.333 3.37e-05 ***
p3            -4.3794     0.9661  -4.533  0.00269 **
p4             5.5627     1.2445   4.470  0.00290 **
p5            -7.3150     1.6257  -4.499  0.00280 **
p6           -23.5477     2.1346 -11.031 1.12e-05 ***
p8           -13.1067     3.9494  -3.319  0.01279 *
Residual standard error: 7.426 on 7 degrees of freedom
Multiple R-Squared: 0.9757, Adjusted R-squared: 0.9548
F-statistic: 46.76 on 6 and 7 DF,  p-value: 2.68e-05
```

This model fits well, but not as well as the best models found by the exhaustive and stepwise methods.

Only for bagpipe (**14**) are all of the model variables are known and YEAR unknown. Using this model to predict YEAR for this pipe results in a 95% prediction interval:

$$\text{Bagpipe } \mathbf{14} \rightarrow (1906.653, 1926.162, 1945.671)\ .$$

This prediction differs from the predictions of the other models we have seen so far.

7. Complete Bagpipe Models

The YEAR, STOCKID and all of the drone variables are known for 13 bagpipes (**1**, **2**, **3**, **4**, **5**, **6**, **9**, **10**, **11**, **12**, **13**, **15**, **18**).

Exhaustive Search.

The best model found by exhaustive search of the 52461 models of rank less than or equal to 3 is

```
YEAR ~ dnew + BMSTCID:BTBUID + BMS2CID:BTTCID
```

This model had both the minimum MSPE (93.27) and AIC (62.53). Among all complete-bagpipe models of rank less than or equal to three, that with the lowest MSPE and AIC uses no information from the tenor drones! This model is the best model found by exhaustive search in Section 6, with the variable dant removed, so no estimation of coefficients or prediction will be done.

Stepwise Search.

A stepwise search algorithm was started at each of the ten best models constructed by the exhaustive method described above, ranked by MSPE. These models are listed in Appendix 3.

Of the ten initial models, three produce search paths which converge to models with finite AIC. All three limit models with finite AIC are different: the one with minimum AIC is the rank 8 model

```
YEAR ~ TTTCID + dnew + dant + TFSID:TTBUID + TT2CID:BTBUID +
               BMS2CID:BTTCID + BMSTCID:BTBUID + BTTCID:BT2CID
```

This model has MSPE = 28.7 and AIC = 51.88. It combines information from the tenor and bass drones not only by using main effects from each, but by including a mixed interaction term, TT2CID:BTBUID. Also, this model does not use STOCKID or BFSID. In R format, the estimated coefficients of this model are

```
fit_lm(YEAR~TTTCID+dnew+dant+TFSID:TTBUID+TT2CID:BTBUID+BMS2CID:BTTCID+
                BMSTCID:BTBUID+BTTCID:BT2CID,x=T)
summary(fit)
                 Estimate Std. Error t value Pr(>abs(t))
(Intercept)     2.224e+03  9.945e+01  22.365 3.32e-06 ***
TTTCID          1.319e-01  1.306e-01   1.010  0.35865
dnew            6.945e+01  5.728e+00  12.125 6.74e-05 ***
dant           -3.639e+01  8.744e+00  -4.162  0.00881 **
TFSID:TTBUID   -5.283e-04  1.598e-04  -3.305  0.02135 *
TT2CID:BTBUID   3.575e-04  1.756e-04   2.036  0.09741 .
BTBUID:BMSTCID  2.394e-05  1.476e-04   0.162  0.87745
BMS2CID:BTTCID -8.965e-04  2.077e-04  -4.317  0.00759 **
BTTCID:BT2CID  -3.765e-04  1.682e-04  -2.238  0.07541 .
Residual standard error: 5.612 on 5 degrees of freedom
Multiple R-Squared: 0.9896, Adjusted R-squared: 0.9729
F-statistic: 59.41 on 8 and 5 DF,  p-value: 0.0001561
```

The residual plots for this model look fair, though the residuals are larger for the three new bagpipes (**10**, **15**, **20**) to which the model can be applied (BMS2CID is unknown for (**22**), the Gibson bagpipe). Perhaps interaction between the dummy variables for approximate age and some continuous predictors should be allowed.

Only for bagpipe (**14**) are all of the model variables are known and YEAR unknown. Using this model to predict YEAR for this pipe results in a 95% prediction interval:

$$\text{Bagpipe } \mathbf{14} \rightarrow (1917.866, 1934.418, 1950.970)\ .$$

Principal Components.

The best model found using principal components is

```
y ~ p[,1] + p[,2] + p[,5] + p[,6] + p[,7] + p[,8]
```

where `p[,i]` is the i^{th} principal component (the loadings of these components have no clear interpretation: see Appendix 4). This model has MSPE = 220.98 and AIC = 76.30. In R format, the estimated coefficients of this model are

```
fit2_lm(y~p[,1]+p[,2]+p[,5]+p[,6]+p[,7]+p[,8],x=T)
summary(fit2)
Coefficients:
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1939.8462     4.4830 432.712 1.02e-14 ***
p[, 1]         3.4558     0.7585   4.556  0.00387 **
p[, 2]        -2.2641     1.3237  -1.710  0.13802
p[, 5]        -3.0505     2.2723  -1.342  0.22802
p[, 6]         6.9175     2.6650   2.596  0.04090 *
p[, 7]        -7.8553     4.2799  -1.835  0.11612
p[, 8]       -15.5158     4.8396  -3.206  0.01846 *
Residual standard error: 16.16 on 6 degrees of freedom
Multiple R-Squared: 0.8843, Adjusted R-squared: 0.7687
F-statistic: 7.645 on 6 and 6 DF,  p-value: 0.01292
```

The quality of fit of this model is disappointing, compared with the best models found by the exhaustive and stepwise methods.

8. Conclusions and Future Work

Conclusions.

We have constructed reasonable-looking linear prediction models for YEAR based on either a tenor drone, a bass drone, or a complete bagpipe. These models were constructed by three different methods, and the exhaustive and stepwise methods applied to the original variables produced better models than the stepwise method applied to models based on principal components.

The best models constructed are fairly consistent in predicting the year of production of the Henderson bagpipe (**14**), as shown by the 95% prediction intervals in Figure 8.1. Clearly, more data is required if we are to test prediction models rigorously.

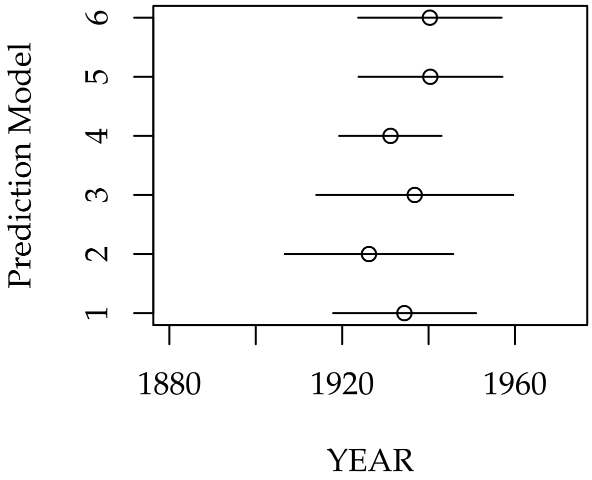

FIGURE 8.1. Six 95% prediction intervals for the Henderson bagpipe (**14**) with unknown YEAR. Each interval was derived from a model in Section 5, 6 or 7. The models are, from top to bottom, tenor drone (exhaustive), tenor drone (stepwise), bass drone (exhaustive), bass drone (stepwise), bass drone (stepwise on principal components) and complete bagpipe (stepwise). The horizonal axis corresponds to Henderson's years in business.

Future Work.

Consider the following scenario: Piper Bob acquires a bagpipe of unknown make and age. Bob would like to know both of these things, and no expert is available to give an opinion (or Bob seeks confirmation of an opinion). Bob measures the internal diameters of a few drone chambers, goes to `www.bagpipe_identification.org` and types the measurements into fields on a webpage. A couple of seconds later, the webpage responds "I'm sure you have a Hardie, and 95% confident it was made between 1958 and 1963" or "this pipe was made between 1910 and 1915 by either Henderson or Lawrie, but the stocks appear to be from about 1980" or "the tenor drones are MacDougalls from about 1875; I cannot identify the bass drone". In the last case, maybe Piper Bob takes a closer look at the bass drone and notices that none of the three sections match exactly.

A web-based tool of this kind clearly would be useful. Judging from the work described in this paper, the author believes that such a tool could be developed using modern classification methods.

References

1. Jeannie Campbell, *Highland Bagpipe Makers*, Magnus Orr Publishing, 2001.
2. Roderick MacLellan, *personal communication.*
3. Roderick MacLellan, *http://www.highland-pipemaker.com/CoverBP.jpg.*

Appendix 1 — The Data

The Data. Roderick MacLellan's measurements of 23 bagpipes.

Appendix 2 — Exploratory Data Analysis

The search for simple linear regression models to predict YEAR produced the following models, which are presented in the form of R output. No other simple models with significant coefficients were found.

The variable names used below are consistent with those in the body of this report. Additionally, `vnew` = (**10**,**15**,**20**,**22**) and `vold` = (**2**,**19**) record the positions in the data set of the new (post-1980) and old (pre-1900) bagpipes: `dnew` and `dold` are dummy variables for these two subsets of the data.

STOCKID.

```
summary(fit_lm(YEAR~STOCKID))
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2559.9651   232.8296   10.99 2.84e-07 ***
STOCKID       -0.7892     0.2923   -2.70   0.0207 *
Residual standard error: 25.5 on 11 degrees of freedom
Multiple R-Squared: 0.3986, Adjusted R-squared: 0.3439
F-statistic:  7.29 on 1 and 11 DF,  p-value: 0.02066
```

TFSID.

```
summary(fit_lm(YEAR~TFSID))
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2259.1319   128.2198  17.619 1.96e-11 ***
TFSID         -0.9754     0.3946  -2.472   0.0259 *
Residual standard error: 30.72 on 15 degrees of freedom
Multiple R-Squared: 0.2895, Adjusted R-squared: 0.2421
F-statistic: 6.111 on 1 and 15 DF,  p-value: 0.0259
```

The Sinclair bagpipe (**3**) has a very large Cook's distance, relative to the the other bagpipes. It can be seen on the extreme left of the scatterplot, well removed from the other data points. The fit of a simple linear model is improved by removing the Sinclair pipe, as shown below.

```
summary(fit_lm(YEAR~TFSID,data=data[-3,]))
            Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2520.283    167.670  15.031 4.95e-10 ***
TFSID         -1.764      0.511  -3.453  0.00388 **
Residual standard error: 27.59 on 14 degrees of freedom
Multiple R-Squared: 0.4599, Adjusted R-squared: 0.4213
F-statistic: 11.92 on 1 and 14 DF,  p-value: 0.003884
```

BMSTCID.

```
summary(fit_lm(YEAR~BMSTCID))
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2274.6406   214.2000  10.619 2.25e-08 ***
BMSTCID       -0.4711     0.3038  -1.551    0.142
Residual standard error: 33.83 on 15 degrees of freedom
Multiple R-Squared: 0.1382, Adjusted R-squared: 0.08072
F-statistic: 2.405 on 1 and 15 DF,  p-value: 0.1418
```

If the antique and new bagpipes are included in the data, then regressing YEAR on BMSTCID does not produce a good fit. Excluding the old and new pipes does produce a good fit, as evident in Figure 3.4 and below.

```
summary(fit_lm(YEAR~BMSTCID,data=data[-c(vnew,vold),]))
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2417.9481   139.0692  17.387 3.11e-08 ***
BMSTCID       -0.6770     0.1955  -3.464  0.00712 **
Residual standard error: 14.08 on 9 degrees of freedom
Multiple R-Squared: 0.5714, Adjusted R-squared: 0.5238
F-statistic:    12 on 1 and 9 DF,  p-value: 0.007117
```

BMS2CID.

```
summary(fit_lm(YEAR~BMS2CID))
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2284.3668   212.3223  10.759 7.65e-08 ***
BMS2CID       -0.8337     0.5125  -1.627    0.128
Residual standard error: 33.66 on 13 degrees of freedom
Multiple R-Squared: 0.1691, Adjusted R-squared: 0.1052
F-statistic: 2.646 on 1 and 13 DF,  p-value: 0.1278
```

If the new bagpipes are included in the data, then regressing YEAR on BMS2CID does not produce a good fit. Excluding the new pipes does produce a good fit, as evident in Figure 3.5 and below.

```
summary(fit_lm(YEAR~BMS2CID,data=data[-c(vnew),]))
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2419.5495   102.8839  23.517 4.38e-10 ***
BMS2CID       -1.1951     0.2493  -4.795  0.00073 ***
Residual standard error: 15.76 on 10 degrees of freedom
Multiple R-Squared: 0.6969, Adjusted R-squared: 0.6665
F-statistic: 22.99 on 1 and 10 DF,  p-value: 0.0007296
```

BTTCID.

```
summary(fit_lm(YEAR~BTTCID))
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2106.1467   361.7681   5.822 3.36e-05 ***
BTTCID        -0.2322     0.5138  -0.452    0.658
Residual standard error: 36.19 on 15 degrees of freedom
Multiple R-Squared: 0.01343, Adjusted R-squared: -0.05234
F-statistic: 0.2042 on 1 and 15 DF,  p-value: 0.6578
```

If the new and antique bagpipes are included in the data, then regressing YEAR on BTTCID does not produce a good fit. Excluding the new and antique pipes does produce a good fit, as evident in Figure 3.6 and below.

```
summary(fit_lm(YEAR~BTTCID,data=data[-c(vnew,vold),]))
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 2493.4797   129.5544  19.247 1.27e-08 ***
BTTCID        -0.7935     0.1845  -4.301  0.00199 **
Residual standard error: 12.31 on 9 degrees of freedom
Multiple R-Squared: 0.6727, Adjusted R-squared: 0.6364
F-statistic:  18.5 on 1 and 9 DF,  p-value: 0.001987
```

Appendix 3 — Exhaustive Model Fitting

Data Processing.

The data is read in,

```
data_read.table("Yggdrasill:Desktop Folder:Laboratory:Bagpipe:bagpipe.dat",
                                           header=T,na.strings="*");
attach(data);
```

different measurements for tenor drones are replaced by their average,

```
for(i in 1:23) if(TFSIDO != "NA") TFSID[i]_0.5*(TFSID[i]+TFSIDO[i])
for(i in 1:23) if(TT2CIDO != "NA") TT2CID[i]_0.5*(TT2CID[i]+TT2CIDO[i])
```

dummy variables are coded for new and antique pipes (here, DANT is denoted `dold` and DNEW is denoted `dnew`),

```
vnew_c(10,15,20,22)
vold_c(2,19)
dnew_c(0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,1,0)
dold_c(0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0)
```

and MAKER, YEAR and the eleven predictors of interest are selected,

```
cdata_data[,c(3,5,7,8,12,14,18,19,22,24,26,28,31)]
```

Tenor Drone Models.

The design matrix X is created to contain STOCKID, the four tenor main effects, the pairwise interactions of these variables, and the dummy variables DNEW and DANT. The corresponding response vector y is also created.

```
X_cbind(cdata[c(3,4,5,6,7)],dnew,dold)
y_cdata[,2]
inter_NULL
b_length(X[1,])-2
idx_1;
for(i in 1:(b-1)) for(j in (i+1):b) {
inter_as.data.frame(cbind(inter,X[,i]*X[,j]))
names(inter)[idx]_paste(names(X)[i],":",names(X)[j]);
idx_idx+1;
}
X_as.data.frame(cbind(X,inter))
X_X[c(1,2,3,4,5,6,9,10,11,12,13,15,18),]
y_y[c(1,2,3,4,5,6,9,10,11,12,13,15,18)]
```

The design matrix contains 17 variables. An exhaustive rountine (written in R by the author) is called to test all models of rank 6 or less based on X and y.

```
tmp_fit.exhaust(X,y,6)
"Testing 21776 models -- approx. run time 0 : 17 : 24"
"Top 10 models in terms of AIC are:"
      V1 V2 V3 V4 V5 V6   AIC  PRESS  MSPE
10055  1  2  5  6  7 11 50.44 465.39 35.80
20215  5  6  8 10 11 17 50.78 516.20 39.71
12485  1  5  6  7  8 11 51.02 473.13 36.39
15600  2  5  6 10 11 17 51.02 548.19 42.17
12509  1  5  6  7 11 14 51.19 486.46 37.42
20333  5  6 10 11 14 17 51.41 552.07 42.47
20330  5  6 10 11 13 17 51.43 572.01 44.00
13109  1  6  8  9 11 16 56.08 817.79 62.91
10364  1  2  6  9 11 16 56.35 845.90 65.07
18885  4  5  6 10 11 13 56.36 646.31 49.72
"Top 10 models in terms of MPSE are:"
      V1 V2 V3 V4 V5 V6   AIC  PRESS  MSPE
10055  1  2  5  6  7 11 50.44 465.39 35.80
12485  1  5  6  7  8 11 51.02 473.13 36.39
12509  1  5  6  7 11 14 51.19 486.46 37.42
20215  5  6  8 10 11 17 50.78 516.20 39.71
15600  2  5  6 10 11 17 51.02 548.19 42.17
20333  5  6 10 11 14 17 51.41 552.07 42.47
20330  5  6 10 11 13 17 51.43 572.01 44.00
18828  4  5  6  8 10 11 56.51 635.84 48.91
18885  4  5  6 10 11 13 56.36 646.31 49.72
18886  4  5  6 10 11 14 56.69 651.99 50.15
```

The best model in terms of both AIC and MSPE is

```
YEAR ~ STOCKID * TTBUID + TFSID + dold + dnew
```

Histograms of the AIC and MSPE for all of the models evaluated are shown in Figure A3.1.

Bass Drone Models.

The design matrix X is created to contain STOCKID, the six bass main effects, the pairwise interactions of these variables, and the dummy variables DNEW and DANT. The corresponding response vector y is also created.

```
X_cbind(cdata[c(3,8,9,10,11,12,13)],dnew,dold)
y_cdata[,2]
inter_NULL
b_length(X[1,])-2
idx_1;
for(i in 1:(b-1)) for(j in (i+1):b) {
inter_as.data.frame(cbind(inter,X[,i]*X[,j]))
names(inter)[idx]_paste(names(X)[i],":",names(X)[j]);
idx_idx+1;
}
X_as.data.frame(cbind(X,inter))
X_X[c(1,2,3,4,5,6,9,10,11,12,13,15,18,19),]
y_y[c(1,2,3,4,5,6,9,10,11,12,13,15,18,19)]
```

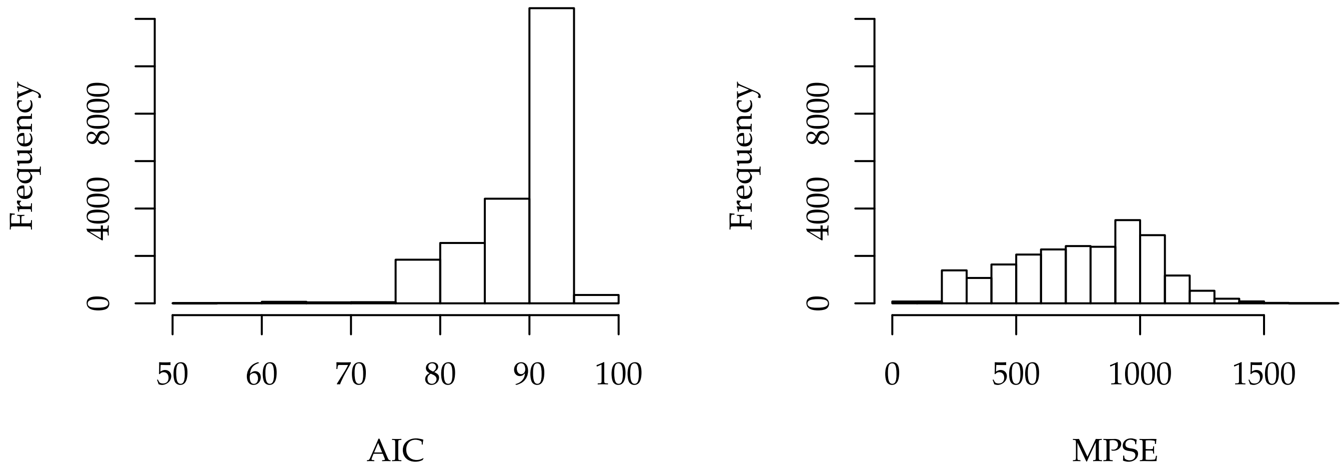

FIGURE A3.1. Histograms of the AIC and MSPE for the 21776 tenor drone models of rank less than or equal to 6.

The design matrix contains 30 variables. An exhaustive rountine is called to test all models of rank 4 or less based on X and y.

```
tmp_fit.exhaust(X,y,4)
"Testing 31929 models -- approx. run time 0 : 25 : 32"
"Top 10 models in terms of AIC are:"
      V1 V2 V3 V4   AIC   PRESS  MSPE
23272  8  9 25 27 61.15  661.49 47.25
21320  7  8  9 25 61.31  683.55 48.83
23265  8  9 24 25 62.46  790.79 56.49
17512  5  8  9 12 63.41 1088.37 77.74
23274  8  9 25 29 63.83  861.80 61.56
23779  8 12 24 25 63.83  977.02 69.79
6619   1  8 24 25 63.91  989.59 70.68
23275  8  9 25 30 64.39  932.60 66.61
24140  8 15 24 25 65.53 1087.82 77.70
15225  4  8  9 25 65.54  996.27 71.16
"Top 10 models in terms of MPSE are:"
      V1 V2 V3 V4   AIC   PRESS  MSPE
23272  8  9 25 27 61.15  661.49 47.25
21320  7  8  9 25 61.31  683.55 48.83
23265  8  9 24 25 62.46  790.79 56.49
23274  8  9 25 29 63.83  861.80 61.56
23275  8  9 25 30 64.39  932.60 66.61
23779  8 12 24 25 63.83  977.02 69.79
6619   1  8 24 25 63.91  989.59 70.68
15225  4  8  9 25 65.54  996.27 71.16
17525  5  8  9 25 65.91 1026.80 73.34
14428  4  5  8  9 66.10 1035.62 73.97
```

The best model in terms of both AIC and MSPE is

```
YEAR ~ dold + dnew + BMSTCID:BTBUID + BMS2CID:BTTCID
```

Histograms of the AIC and MSPE for all of the models evaluated are shown in Figure A3.2.

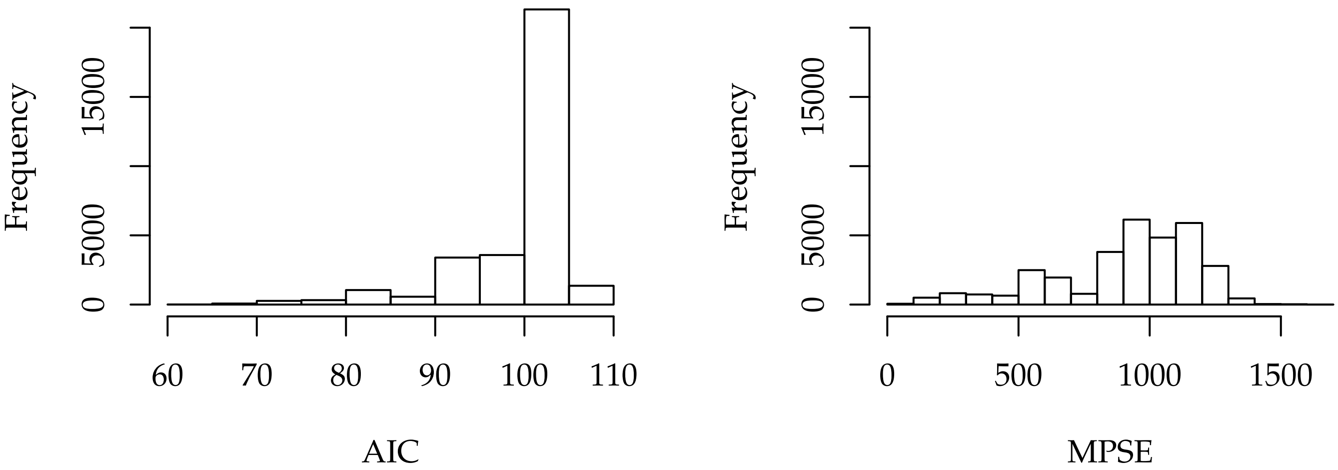

FIGURE A3.1. Histograms of the AIC and MSPE for the 31929 bass drone models of rank less than or equal to 4.

Complete-Bagpipe Models.

The design matrix X is created to contain all eleven main effects, the pairwise interactions of these variables, and the dummy variables DNEW and DANT. The corresponding response vector y is also created.

```
X_cbind(cdata[,-c(1,2)],dnew,dold)
y_cdata[,2]
inter_NULL
b_length(X[1,])-2
idx_1;
for(i in 1:(b-1)) for(j in (i+1):b) {
inter_as.data.frame(cbind(inter,X[,i]*X[,j]))
names(inter)[idx]_paste(names(X)[i],":",names(X)[j]);
idx_idx+1;
}
X_as.data.frame(cbind(X,inter))
X_X[c(1,2,3,4,5,6,9,10,11,12,13,15,18),]
y_y[c(1,2,3,4,5,6,9,10,11,12,13,15,18)]
```

The design matrix contains 68 variables. An exhaustive rountine is called to test all models of rank 3 or less based on X and y.

```
tmp_fit.exhaust(X,y,3)
"Testing 52461 models -- approx. run time 0 : 41 : 57"
"Top 10 models in terms of AIC are:"
      V1 V2 V3   AIC   PRESS   MSPE
24722 12 62 63 62.53 1212.55  93.27
24731 12 63 67 63.73 1299.80  99.98
24527 12 47 63 64.23 1430.16 110.01
24732 12 63 68 64.52 1423.20 109.48
23252 12 13 63 64.55 1338.98 103.00
18358  9 12 13 65.67 1328.91 102.22
24359 12 40 63 65.92 1539.48 118.42
16674  8 12 40 66.08 1382.77 106.37
23219 12 13 30 66.11 1567.76 120.60
21657 11 12 63 66.58 1549.61 119.20
"Top 10 models in terms of MPSE are:"
      V1 V2 V3   AIC   PRESS   MSPE
24722 12 62 63 62.53 1212.55  93.27
24731 12 63 67 63.73 1299.80  99.98
18358  9 12 13 65.67 1328.91 102.22
23252 12 13 63 64.55 1338.98 103.00
16674  8 12 40 66.08 1382.77 106.37
24732 12 63 68 64.52 1423.20 109.48
24527 12 47 63 64.23 1430.16 110.01
23249 12 13 60 67.23 1493.60 114.89
24359 12 40 63 65.92 1539.48 118.42
21657 11 12 63 66.58 1549.61 119.20
```

The best model in terms of both AIC and MSPE is

```
YEAR ~ dnew + BMSTCID:BTBUID + BMS2CID:BTTCID
```

Histograms of the AIC and MSPE for all of the models evaluated are shown in Figure A3.3.

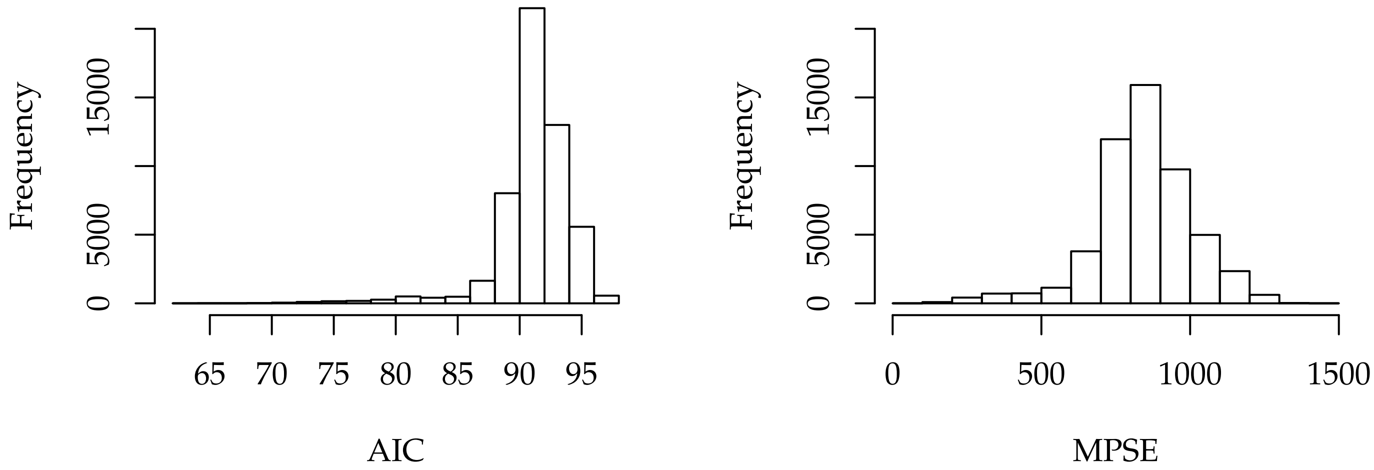

FIGURE A3.3. Histograms of the AIC and MSPE for the 52461 complete-bagpipe models of rank less than or equal to 3.

.

Tenor Drone Models.

The design matrix X is created just as in Appendix 3: it contains 17 variables. Principal components are computed using the `princomp` function in the "mva" R library,

```
pc.cor_princomp(X,cor=T)
summary(pc.cor)
                          Comp.1    Comp.2    Comp.3    Comp.4
Standard deviation     2.9448397 1.6512076 1.4206578 1.3355272
Proportion of Variance 0.5101224 0.1603816 0.1187217 0.1049196
Cumulative Proportion  0.5101224 0.6705040 0.7892257 0.8941452
                           Comp.5     Comp.6     Comp.7
Standard deviation     0.96394707 0.76693324 0.53071967
Proportion of Variance 0.05465847 0.03459921 0.01656843
Cumulative Proportion  0.94880371 0.98340292 0.99997135
```

The first seven principal components capture 99.997% of the variance in the data. The corresponding scree plot is shown in Figure A4.1.

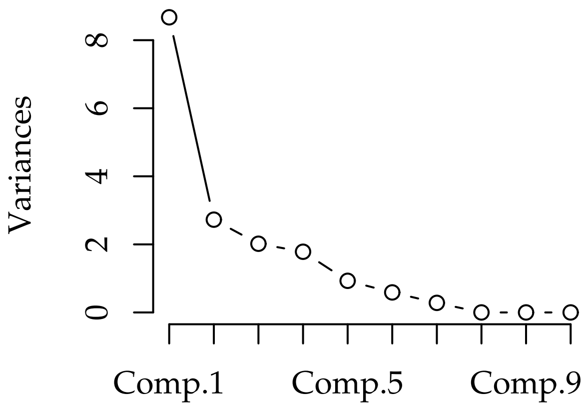

Figure A4.1. Scree plot for the pricipal component analysis of the tenor drone data.

The loadings for the first seven principal components are

```
                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
STOCKID          -0.277  0.172  0.224         0.323 -0.216  0.320
TFSID            -0.258  0.338 -0.185        -0.137  0.169
TTTCID           -0.135 -0.389         0.446 -0.228  0.135  0.242
TT2CID           -0.117 -0.246 -0.535 -0.177        -0.334 -0.169
TTBUID           -0.237 -0.203  0.290 -0.331         0.178 -0.195
dnew              0.103               -0.457 -0.649 -0.238  0.539
dold             -0.119  0.110  0.334  0.318 -0.381 -0.651 -0.437
STOCKID : TFSID  -0.291  0.304
STOCKID : TTTCID -0.295         0.164  0.249  0.102         0.395
STOCKID : TT2CID -0.295        -0.130 -0.124  0.325 -0.364  0.162
STOCKID : TTBUID -0.289         0.295 -0.220  0.105
TFSID : TTTCID   -0.285  0.184 -0.171  0.207 -0.209  0.210
TFSID : TT2CID   -0.268  0.230 -0.330                       -0.119
TFSID : TTBUID   -0.314  0.161        -0.106 -0.152  0.207 -0.143
TTTCID : TT2CID  -0.161 -0.414 -0.334  0.193        -0.104
TTTCID : TTBUID  -0.249 -0.341  0.215        -0.176  0.201
TT2CID : TTBUID  -0.250 -0.281        -0.355                -0.241
```

No meaningful interpretation of the principal components is evident from the loadings. Fitting the data with the first seven principal components results in a model with MSPE = 416.94 and AIC = 80.42. In R format, the estimated coefficients of this model are

```
p_pc.cor$scores
fit1_lm(y~p[,1]+p[,2]+p[,3]+p[,4]+p[,5]+p[,6]+p[,7],x=T)
summary(fit)
              Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1939.84615    5.32818 364.073 2.97e-12 ***
p[, 1]         7.03916    1.80933   3.890   0.0115 *
p[, 2]        -2.02256    3.22684  -0.627   0.5583
p[, 3]        -2.00744    3.75050  -0.535   0.6154
p[, 4]       -12.61082    3.98957  -3.161   0.0251 *
p[, 5]        -5.27397    5.52746  -0.954   0.3838
p[, 6]         0.01576    6.94738   0.002   0.9983
p[, 7]        22.45343   10.03954   2.236   0.0755 .
Residual standard error: 19.21 on 5 degrees of freedom
Multiple R-Squared: 0.8638, Adjusted R-squared: 0.6732
F-statistic: 4.531 on 7 and 5 DF,  p-value: 0.05767
```

This is not a particularly good fit. From the t-statistics, it looks as if only the first, fourth and seventh principal components matter. Staring at the model `fit1` above, stepwise model selection (restricted to the first seven principal components) converges to

```
y ~ p[,1] + p[,4] + p[,7]
```

which has MSPE = 241.25 and AIC = 76.01. In R format, the estimated coefficients of this model are

```
fit2_lm(y~p[,1]+p[,4]+p[,7],x=T)
summary(fit2)
            Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1939.846      4.559 425.476  < 2e-16 ***
p[, 1]         7.039      1.548   4.547  0.00139 **
p[, 4]       -12.611      3.414  -3.694  0.00497 **
p[, 7]        22.453      8.591   2.614  0.02810 *
Residual standard error: 16.44 on 9 degrees of freedom
Multiple R-Squared: 0.8205, Adjusted R-squared: 0.7607
F-statistic: 13.72 on 3 and 9 DF,  p-value: 0.00105
```

Bass Drone Models.

The design matrix X is created just as in Appendix 3: it contains 30 variables. Principal components are computed using the `princomp` function in the R "mva" library,

```
pc.cor_princomp(X,cor=T)
summary(pc.cor)
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     3.7959868 2.3756882 2.0543363 1.59474087
Proportion of Variance 0.4803172 0.1881298 0.1406766 0.08477328
Cumulative Proportion  0.4803172 0.6684470 0.8091236 0.89389688
                           Comp.5     Comp.6     Comp.7
Standard deviation     1.22079058 0.92975494 0.70074715
Proportion of Variance 0.04967765 0.02881481 0.01636822
Cumulative Proportion  0.94357454 0.97238935 0.98875757
                            Comp.8      Comp.9      Comp.10
Standard deviation     0.502532906 0.287097280 3.805302e-02
Proportion of Variance 0.008417977 0.002747495 4.826773e-05
Cumulative Proportion  0.997175543 0.999923038 9.999713e-01
```

The first ten principal components capture 99.997% of the variance in the data. The corresponding scree plot is shown in Figure A4.2.

As in the case of the tenor drones, no meaningful interpretation was evident from the loadings. Fitting the data with the first eight principal components results in a model with MSPE = 64.81 and AIC = 61.22. In R format, the estimated coefficients of this model are

```
p_pc.cor$scores
fit1_lm(y~p[,1]+p[,2]+p[,3]+p[,4]+p[,5]+p[,6]+p[,7]+p[,8],x=T)
summary(fit1)
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1936.2857     2.0936 924.853  2.8e-14 ***
p[, 1]         4.8797     0.5515   8.847 0.000307 ***
p[, 2]        -0.6622     0.8813  -0.751 0.486263
p[, 3]        -4.3794     1.0191  -4.297 0.007736 **
p[, 4]         5.5627     1.3128   4.237 0.008191 **
p[, 5]        -7.3150     1.7150  -4.265 0.007974 **
p[, 6]       -23.5477     2.2518 -10.457 0.000138 ***
p[, 7]        -2.5458     2.9877  -0.852 0.433072
p[, 8]       -13.1067     4.1661  -3.146 0.025493 *
Residual standard error: 7.834 on 5 degrees of freedom
Multiple R-Squared: 0.9807, Adjusted R-squared: 0.9497
F-statistic: 31.68 on 8 and 5 DF,  p-value: 0.000721
```

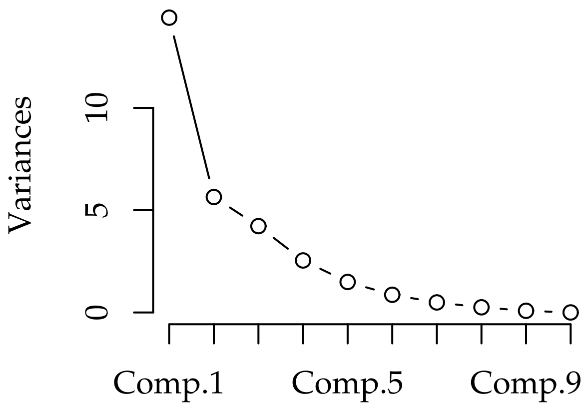

FIGURE A4.2. Scree plot for the pricipal component analysis of the bass drone data.

This is a good fit, but we can do better. From the t-statistics, it looks as if the second and seventh principal components don't matter. Staring at the model `fit1` above, stepwise model selection (restricted to the first eight principal components) converges to

```
y ~ p[,1] + p[,3] + p[,4] + p[,5] + p[,6] + p[,8]
```

which has MSPE = 59.1 and AIC = 60.44. In R format, the estimated coefficients of this model are

```
p1_p[,1]; p3_p[,3]; p4_p[,4]; p5_p[,5]; p6_p[,6]; p8_p[,8];
fit2_lm(y~p1+p3+p4+p5+p6+p8,x=T)
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1936.2857     1.9847 975.607  < 2e-16 ***
p1             4.8797     0.5228   9.333 3.37e-05 ***
p3            -4.3794     0.9661  -4.533  0.00269 **
p4             5.5627     1.2445   4.470  0.00290 **
p5            -7.3150     1.6257  -4.499  0.00280 **
p6           -23.5477     2.1346 -11.031 1.12e-05 ***
p8           -13.1067     3.9494  -3.319  0.01279 *
Residual standard error: 7.426 on 7 degrees of freedom
Multiple R-Squared: 0.9757, Adjusted R-squared: 0.9548
F-statistic: 46.76 on 6 and 7 DF,  p-value: 2.68e-05
```

The loadings of the principal components used in this model are

```
round(pc.cor$loadings[,c(1,3,4,5,6,8)],3)
                  Comp.1 Comp.3 Comp.4 Comp.5 Comp.6 Comp.8
STOCKID           -0.136  0.185 -0.431  0.199 -0.191  0.000
BFSID             -0.134 -0.050  0.030 -0.093  0.004  0.040
BMSTCID           -0.209 -0.242  0.072  0.128  0.087 -0.424
BMS2CID           -0.198  0.242  0.239  0.092  0.061  0.013
BTTCID            -0.171 -0.262  0.157  0.261 -0.023  0.470
BT2CID            -0.054 -0.323 -0.155 -0.172  0.106 -0.053
BTBUID            -0.195  0.173  0.003 -0.329 -0.064  0.010
dnew               0.043 -0.003  0.362  0.174 -0.753 -0.270
dold              -0.030  0.343  0.005  0.340  0.504 -0.153
STOCKID : BFSID   -0.155  0.004 -0.086 -0.032 -0.040  0.020
STOCKID : BMSTCID -0.224 -0.033 -0.245  0.216 -0.077 -0.281
STOCKID : BMS2CID -0.211  0.269 -0.045  0.166 -0.048  0.005
STOCKID : BTTCID  -0.210 -0.009 -0.256  0.314 -0.171  0.273
STOCKID : BT2CID  -0.151 -0.052 -0.468  0.062 -0.106 -0.027
STOCKID : BTBUID  -0.206  0.211 -0.189 -0.153 -0.138  0.015
BFSID : BMSTCID   -0.168 -0.104  0.046 -0.048  0.031 -0.096
BFSID : BMS2CID   -0.185  0.043  0.113 -0.048  0.030  0.047
BFSID : BTTCID    -0.162 -0.103  0.065 -0.032  0.004  0.137
BFSID : BT2CID    -0.151 -0.128 -0.006 -0.140  0.019  0.025
BFSID : BTBUID    -0.201  0.030  0.030 -0.234 -0.026  0.038
BMSTCID : BMS2CID -0.240  0.055  0.203  0.125  0.084 -0.195
BMSTCID : BTTCID  -0.201 -0.264  0.112  0.194  0.042 -0.030
BMSTCID : BT2CID  -0.168 -0.333 -0.038 -0.004  0.113 -0.315
BMSTCID : BTBUID  -0.239  0.023  0.031 -0.194 -0.013 -0.182
BMS2CID : BTTCID  -0.223  0.080  0.250  0.177  0.037  0.209
BMS2CID : BT2CID  -0.218  0.065  0.149 -0.002  0.108 -0.018
BMS2CID : BTBUID  -0.211  0.223  0.116 -0.138 -0.008  0.017
BTTCID : BT2CID   -0.130 -0.342 -0.005  0.051  0.048  0.241
BTTCID : BTBUID   -0.226  0.044  0.057 -0.175 -0.065  0.191
BT2CID : BTBUID   -0.187  0.019 -0.062 -0.347 -0.014 -0.003
```

We can test the predictive ability of this model by putting the measurements for bagpipe (**14**) back into the design matrix X are re-computing the principal components. This results in

```
round(pc.cor$scores[12,],3)
 Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8
 -2.067  -2.083   1.262  -1.249  -0.796  -0.444  -0.350   0.293
 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16
  0.587  -0.010  -0.015   0.015   0.006   0.020   0.000   0.000
Comp.17 Comp.18 Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24
  0.000   0.000   0.000   0.000   0.000   0.000   0.000   0.000
Comp.25 Comp.26 Comp.27 Comp.28 Comp.29 Comp.30
  0.000   0.000   0.000   0.000   0.000   0.000
```

and so we predict by

```
new_data.frame(y=0,p1=-2.067,p3=1.262,p4=-1.249,p5=-0.796,p6=-0.444,p8=0.293)
predict.lm(fit2,new,interval="prediction",level=0.95)
         fit      lwr      upr
[1,] 1926.162 1906.653 1945.671
```

Complete-Bagpipe Models.

The design matrix X is created just as in Appendix 3: it contains 68 variables. Principal components are computed using the `princomp` function in the R "mva" library,

```
pc.cor_princomp(X,cor=T)
summary(pc.cor)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4
Standard deviation     5.9101628 3.3868215 2.7341185 2.1213369
Proportion of Variance 0.5136768 0.1686847 0.1099324 0.0661775
Cumulative Proportion  0.5136768 0.6823615 0.7922939 0.8584714
                           Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     1.97285571 1.68215378 1.04745964 0.92632009
Proportion of Variance 0.05723764 0.04161237 0.01613488 0.01261866
Cumulative Proportion  0.91570908 0.95732145 0.97345633 0.98607499
                           Comp.9     Comp.10     Comp.11
Standard deviation     0.72244625 0.582442852 0.289746444
Proportion of Variance 0.00767542 0.004988819 0.001234603
Cumulative Proportion  0.99375041 0.998739229 0.999973832
```

The first eleven principal components capture 99.997% of the variance in the data. The corresponding scree plot is shown in Figure A4.3.

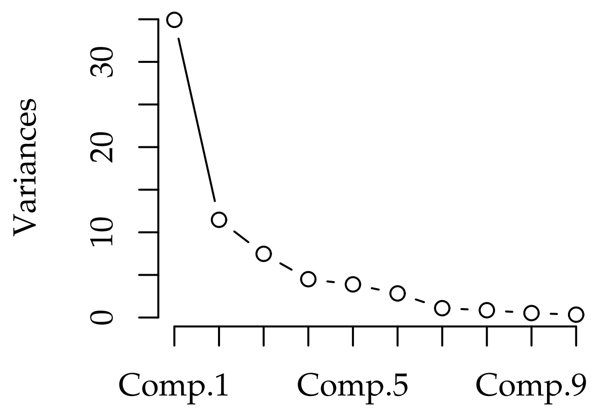

FIGURE A4.3. Scree plot for the pricipal component analysis of the complete-bagpipe data.

No meaningful interpretation was evident from the loadings. Fitting the data with the first eight principal components results in a model with MSPE = 416.94 and AIC = 80.42. In R format, the estimated coefficients of this model are

```
p_pc.cor$scores
fit1_lm(y~p[,1]+p[,2]+p[,3]+p[,4]+p[,5]+p[,6]+p[,7]+p[,8],x=T)
summary(fit1)
Coefficients:
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1939.8462     5.4386 356.684 3.71e-10 ***
p[, 1]         3.4558     0.9202   3.756   0.0199 *
p[, 2]        -2.2641     1.6058  -1.410   0.2314
p[, 3]         0.5311     1.9891   0.267   0.8027
p[, 4]         0.1905     2.5637   0.074   0.9443
p[, 5]        -3.0505     2.7567  -1.107   0.3305
p[, 6]         6.9175     3.2331   2.140   0.0991 .
p[, 7]        -7.8553     5.1921  -1.513   0.2049
p[, 8]       -15.5158     5.8711  -2.643   0.0574 .
Residual standard error: 19.61 on 4 degrees of freedom
Multiple R-Squared: 0.8865, Adjusted R-squared: 0.6595
F-statistic: 3.905 on 8 and 4 DF,  p-value: 0.102
```

This is not a very good fit. From the t-statistics, it looks as if only the first, sixth and eighth principal components matter. Staring at the model `fit1` above, stepwise model selection (restricted to the first eight principal components) converges to

```
y ~ p[,1] + p[,2] + p[,5] + p[,6] + p[,7] + p[,8]
```

which has MSPE = 220.98 and AIC = 76.30. In R format, the estimated coefficients of this model are

```
fit2_lm(y~p[,1]+p[,2]+p[,5]+p[,6]+p[,7]+p[,8],x=T)
summary(fit2)
Coefficients:
             Estimate Std. Error t value Pr(>abs(t))
(Intercept) 1939.8462     4.4830 432.712 1.02e-14 ***
p[, 1]         3.4558     0.7585   4.556  0.00387 **
p[, 2]        -2.2641     1.3237  -1.710  0.13802
p[, 5]        -3.0505     2.2723  -1.342  0.22802
p[, 6]         6.9175     2.6650   2.596  0.04090 *
p[, 7]        -7.8553     4.2799  -1.835  0.11612
p[, 8]       -15.5158     4.8396  -3.206  0.01846 *
Residual standard error: 16.16 on 6 degrees of freedom
Multiple R-Squared: 0.8843, Adjusted R-squared: 0.7687
F-statistic: 7.645 on 6 and 6 DF,  p-value: 0.01292
```