Bare Hull Upright Resistance Prediction Based on the Delft Systematic Yacht Hull Series

Abstract. A regression analysis yields a new formula for the prediction of the resistance of the models of the DSYHS that seems to be superior to the original Delft-method. Future comparisons with more test results have to tell if the new formula is also better when applied to other hulls, not considered in the regression.

NOMENCLATURE

A_{wet}	Wetted area of canoe body
A_W	Water plane area
A_X	Area of max. section = maximum immersed area of all cross sections
BM	Longitudinal metacentric height
B_{WL}	Maximum beam in the water plane
B_X	Beam in the water plane at max. section
C_P	Prismatic coefficient
C_{WP}	Waterplane area coefficient
C_X	Maximum section coefficient
D	equivalent diameter for body of revolution
E_R	Exit angle of the rocker at aft end L_{WL}
Fn	Froude number $U_\infty / (g \cdot L_{WL})^{1/2}$
g	gravitational acceleration = 9.81 m/s^2
G	wetted girth-length of the cross section
I_E	Incidence angle at water plane entrance
L_1	Distance from front of L_{WL} to max. section
L_2	Distance from max. section to aft end L_{WL}
Lcb	Distance front of L_{WL} to centre of buoyancy
Lcf	Distance front of L_{WL} to centre of flotation
L_{TO}	Length of transom overhang
L_{WL}	Length of the actual water line at rest with trimming moment applied
L_{WL0}	Length of the designed water line
r	Pearson's correlation coefficient
R	Resistance force
$\mathscr{R}^2$	Coefficient of determination = r^2
Re	Reynolds number $U_\infty \cdot L_{WL} / \nu$
Tu	Turbulence level
T_{CB}	Maximum draft of canoe body
T_X	Draft of canoe body at max. section
U	Water speed at edge of boundary layer
U_∞	Ship speed
u',v',w'	Fluctuating velocities
V_{attn}	Volume attenuated with depth
V_{CB}	Displaced volume of canoe body
V_1	Part of V_{CB} forward of max. section
V_2	Part of V_{CB} aft of max. section
x,y,z	Hull coordinates
α	Flare angle at max. section at L_{WL}
ρ	Density of the water
ν	Kinematic viscosity of the water
σ	Standard deviation

1. INTRODUCTION

Most velocity prediction programs (VPP) rely on the regression formulas of the Delft Systematic Yacht Hull Series (DSYHS) [1] for the estimation of the bare hull resistance. The basic method of the analysis of the towing tank data was already established in 1973 by Professor Gerritsma. An estimate of the viscous resistance was subtracted from the measured total resistance of the model. The remaining difference, the residuary resistance, was correlated to the geometric properties of the hull. Since computer power has dramatically increased in the last 40 years, there are more sophisticated ways today to predict the viscous resistance than the simple methods of 1973. The aim of this paper is a critical review of the Delft-method and an investigation of other correlation methods for the hull resistance, including the results of the newest research in statistics, with a hopefully better goodness-of-fit to the data. This new approach is only possible because the researchers at Delft University of Technology decided in a really generous act to make the entire experimental data sets of all tested hulls available to the public. Since the beginning of 2013 the data is online [2].

There are more test results of systematic yacht hull series available in the literature, but none of them is as complete as the experimental data from Delft. Battistin et al. [3] published the results of systematic tests of IACC-yachts at INSEAN. All the relevant data is supplied but with the exception of the varying trimming moment that was applied during each run. With this vital information missing, these experiments can unfortunately not be used for a regression analysis.

Interesting work is reported by Huetz and Guillerm [4], who created the database for the regression not from towing tank experiments but by CFD. They used the RANSE free-surface solver ICARE and claim to get a more reliable database with CFD because they avoid the measurement errors of the towing tank. The drawback is the high computing effort that only allowed the simulation at three Froude numbers, namely 0.35, 0.5 and 0.65. Huetz and Guillerm checked the validity of the results with three models out of the DSYHS; a diagram is shown for Sysser 25. In this diagram the total drag values from CFD and towing tank agree within 5% with the

exception of *Fn* = 0.35, where the CFD-prediction is 20% too high. Their published coefficients were used by this author to predict the residuary resistance for Sysser 71, 72 and 73, because these hull forms are close to the ones used in [4]. Surprisingly the errors were larger than 50%. A similar experience was made by van Mierlo [5] who compared the residuary resistance of 38 models of the DSYHS with CFD-results computed with SHIPFLOW. Even though he worked together with the developers of the code at Gothenburg, he was not able to explain the large discrepancies between measurements and predictions. At *Fn* = 0.35 e.g. he found a deviation of ± 50%. At Delft a repeatability of 2% is claimed for the test results [6]. At this time it is not clear whether CFD-results are accurate enough to form a database for a regression analysis, but progress is evident [36].

2. RESISTANCE COMPONENTS OF THE HULL

In his first step Gerritsma followed Froude's method. Froude's work, published in 1868, marked the beginning of scientific ship model testing. The towing tank became the most efficient design tool in ship optimization. Details can be found in any text book on naval architecture e.g. [7]. The basic idea is to split the measured total resistance of the model into a viscous part, which scales with the Reynolds number, and into a wave making part, which scales with the Froude number. The extrapolated values at full scale are then added to give the total resistance of the full size ship. The determination of the viscous resistance was the subject of intense debates over decades during the meetings of the International Towing Tank Conference. It is not possible to calculate or measure the true viscous resistance of the model, including the effects of sinkage and trim and of the varying wave height along the ship; instead the value is determined for the still water level. The remaining resistance component, after the subtraction of the viscous estimate, contained therefore not only the wave making resistance but also some small viscous "left overs". This remainder was therefore not called wave resistance but more appropriately residuary resistance. Because of the small viscous effects that are contained in the residuary resistance, Froude scaling to full size is never 100% correct.

The Delft-method is based on the ITTC procedure of 1957. The viscous drag according to the ITTC is computed from:

$$R_{visc} = \tfrac{1}{2} \cdot \rho \cdot U_\infty{}^2 \cdot C_{ITTC} \cdot (1+k) \cdot A_{wet} \tag{1}$$

C_{ITTC} is the friction coefficient of the ITTC-57 correlation line and k is the form factor determined experimentally by Prohaskas's method. It is questionable if the ITTC-57 correlation line will give good results for yacht like bodies. It is based on the skin friction of flat plates and was developed for merchant marine vessels with long parallel sides that do not exist in yachts. The Delft-method tries to compensate for that by using only 70% of L_{WL0} in the determination of the Reynolds number. The modified friction coefficient used in the Delft-formula that replaces C_{ITTC} reads:

$$C_{Delft} = \frac{0.075}{(\log_{10} Re - 2)^2} \quad \text{with} \quad Re = \frac{U_\infty \cdot 0.7 \cdot L_{WL0}}{\nu} \tag{2}$$

The Delft-method does not use a form factor. The viscous form drag is therefore part of the residuary resistance. This does not pose a problem for the correlation of the data at model size, but it will introduce an error when extrapolating to full size, since viscous drag and residuary drag are scaled differently.

With today's computing power there is no need any more to rely on the ITTC-57 correlation line. Even in fast VPPs it is possible to include a simple boundary layer calculation that determines the viscous resistance with a higher accuracy than the correlation-line-methods. A three-dimensional integral boundary layer calculation method for bodies of revolution is described in [8]. To make use of this fast method, the hull is approximated by half a body of revolution. The equivalent diameter for the body of revolution can be determined in many different ways. A comparison with all the models of the Delft-series gave the best results for a combination that is formally identical to the definition of the hydraulic diameter in pipe flow. For each cross section with an immersed area A, the equivalent diameter is calculated from:

$$D = \frac{4 \cdot A}{G} \tag{3}$$

The velocity at the edge of the boundary layer is taken from the potential flow solution [9]. The differential equations for the boundary layer are then solved by the scheme of Stoer and Bulirsch. The solution of v. Kármán's momentum equation yields the local skin friction coefficient c_f (the dimensionless form of the shear stress) and the b.l.-momentum thickness as a function of the position along the hull. Figure 1 compares the calculated local skin friction coefficient along the hull with that of flat plates of equivalent length and different

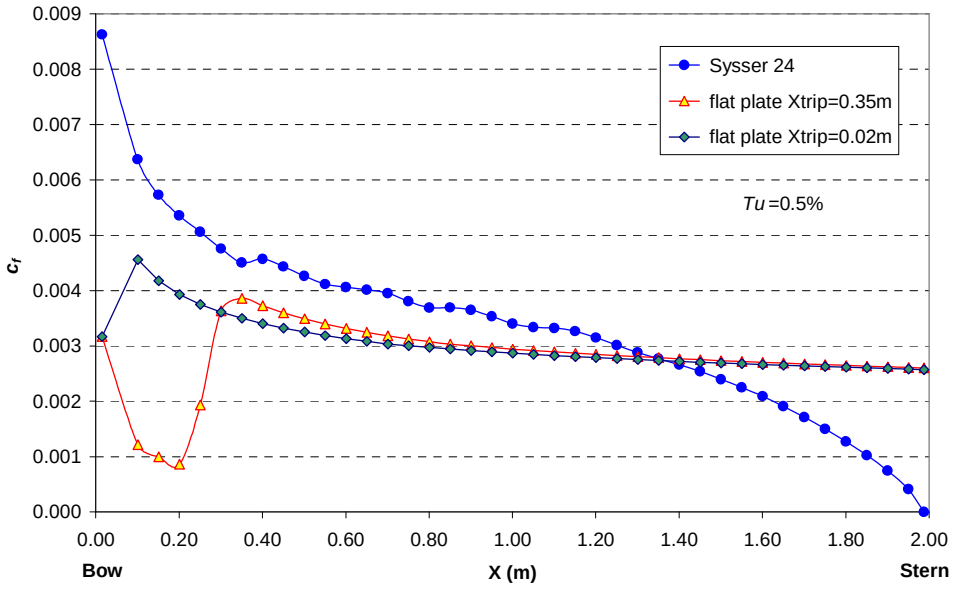

Figure 1. Local skin friction coefficient at 1.77 m/s ship speed

positions of the trip. There are two active boundary layer trips on the model, the streamlines close to the keel are tripped 0.02m behind the bow and the streamlines at the sides 0.35m behind the bow. On the flat plate with the trip at 0.02m the trip is active whereas on the flat plate with the trip at 0.35m it is not active, because along the plate there is no flow acceleration and natural transition occurs at 0.2m. It is obvious that the local skin friction on a flat plate is significantly different from the development of c_f along a ship's hull. In the ITTC procedure the burden to correct this flat plate drag in such a way that the result agrees with the hull drag is put entirely on the form factor k. It is not surprising, that no simple method exists to predict the value of this form factor. The problem increases if b.l.-trips are used. The experiment gives no indication, which one of the trips is causing the transition. At Delft they assumed turbulent flow, starting right at the bow.

The implemented integral computing method performs an integration of the local shear forces from bow to stern which yields the total skin friction drag according to:

$$R_f = \int_0^{L_{WL}} \tfrac{1}{2}\rho \cdot U^2 \cdot G \cdot c_f \cdot \mathrm{d}\,x \tag{4}$$

As this is only the skin friction drag, the addition of the pressure drag is needed to give the total viscous drag. When an integral b.l.-calculation is employed the total drag is usually calculated directly from the momentum deficit thickness at the tail of the body, using either Squire-Young's or Granvilles's [10] formula, which make use of v. Kármán's momentum equation. In our case with a very thick b.l. at the stern, that is often separated, the application of this formula results in unrealistic values for the total drag. v. Kármán's momentum equation is only applicable if the pressure at right angles to the wall is constant within the b.l. This is a realistic assumption for a thin b.l., but in the very thick b.l. at the stern with strongly curved streamlines this assumption is violated. Patel and Guven proved this experimentally [11]. In our case it is therefore preferable to rely on the very robust integration of equation 4 and add to this an estimate for the pressure drag. Hoerner [12] proposes for streamlined bodies of revolution a form factor that consists of two terms. He explains that his first term takes care of the drag increase due to the supervelocity along the curved streamlines and the second term simulates the pressure drag due to the thickness of the boundary layer at the tail that reduces the curvature of the streamlines and therefore the pressure recovery. This second term is a good estimate for the required pressure drag, in case the flow is attached. The viscous drag is therefore calculated from:

$$R_{visc} = R_f \cdot \left(1 + 7 \cdot \left(\frac{D_{max}}{2 \cdot L_2}\right)^3\right) \tag{5}$$

Should the b.l.-calculation detect flow separation, Hoerner's formula for the base drag is used instead to calculate the pressure drag. This method was successfully validated against test data from the wind tunnel and also with CFD-results of a RANS-solver.

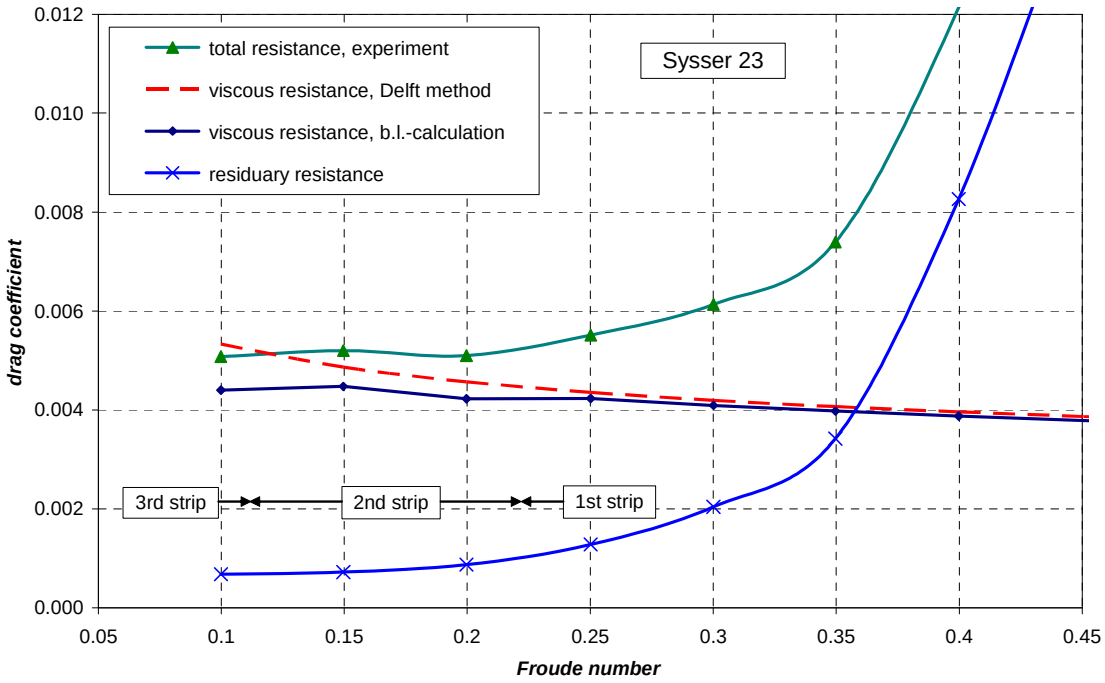

Figure 2. Calculated and measured drag coefficients for Sysser 23

The comparison of the Delft-method and the b.l. calculation is depicted in figure 2. Sysser 23 is a good example that shows the influence of the sand strips and how the Delft-method (equation 2) overpredicts the viscous drag at low speeds. The drag coefficient is here defined as:

$$C = \frac{R}{\frac{1}{2} \cdot \rho \cdot U_\infty{}^2 \cdot A_{wet}} \tag{6}$$

The viscous resistance calculated by the Delft-method is definitely too high at Fn = 0.1; it is even higher than the measured total resistance, which is the sum of viscous and wave drag. The reason for the inadequacy of the turbulent friction coefficient is the large extent of laminar flow at low speeds with these small models. At Delft, they tried to eliminate the laminar flow by using b.l.-trips in the form of carborundum sand strips, applied vertically to the hull surface at three fixed distances from the bow. Each of the three strips was 20 mm wide and the distance between the strips was 240 mm on the small models and 330 mm on the large ones [13]. Whether one of these three strips is capable to trip the flow from laminar to turbulent depends on the Reynolds number, based on the momentum thickness of the b.l. A prediction model for this type of flow is described in [14]. The viscous resistance curve from the integral b.l. calculation shows, that the transition point is jumping from the 3rd roughness strip to the 2nd at a Froude number above 0.1. A second jump occurs between 0.2 and 0.25, when the transition point moves from the 2nd to the 1st strip at the bow. These sudden increases in the viscous resistance are also visible in the total measured resistance, whereas the curvature of the residuary resistance is smooth. The residuary resistance as a function of the Froude number for all the models from no. 1 to no. 73 was determined with this method.

An input parameter to the b.l. calculation is the turbulence level in the flow around the hull. It is defined as:

$$Tu = \frac{\sqrt{\left(\overline{u'^2} + \overline{v'^2} + \overline{w'^2}\right)/3}}{U_\infty} \tag{7}$$

The fluctuating velocities are created by two different mechanisms. One is the breaking of the bow wave. The vorticity in the surface layer of the bow wave can increase Tu up to a value of several percent [15], but it is dissipated quickly and does not reach the lower parts of the hull. To account for the bow wave effect, a minimum value of Tu = 0.15% is prescribed in the calculation. The other root cause of the fluctuating velocities are eddies in the wake of the towed model that persist in the tank a long time after the run. These eddies finally dissipate, but the waiting time between the runs determines obviously the turbulence level in the tank. If the rotating speeds of the eddies are regarded as constant during the run, then Tu will decrease as the towing speed U_∞ increases. The importance of Tu is well known to practitioners for a long time. In his comment at the 6th ITTC in 1951 Mr. Ferguson from Clydebank explained [16]:

> We tried that {i.e. a trip wire}, and certainly got an increase in measured resistance, but the values were not steady. We studied these tests and varied the intervals of time between successive test runs. There was much greater difference between the curves so obtained than between the curves for the naked model and for the model with trip wire. To us it appeared that the quality of the degree of internal turbulence in water was as important as the type of turbulence stimulator applied to the model. The "Monday morning feeling" has been quoted at times. By that I mean the interval over the weekend when the water has been allowed to lie quietly. This affects the first runs you do on Monday morning. We always ignore the first three or four runs we do on Monday morning, until the water gets into what I call a uniform state of consistent turbulence, and after that the results will be comparable. Even during the lunch period and the first run after lunch you do get a difference if you repeat a spot at the same speed.

At Delft they used a waiting time between runs of 15 minutes, which seems to be industry standard. What is unusual is the statement that they never repeated a run during a test campaign [6]. Without these checks it is very difficult to catch outliers. Several outliers at low speeds could be explained and corrected by a variation of *Tu* in the b.l. calculation. For the DSYHS the value $Tu = 1/U_\infty$ (speed in m/s yields *Tu* in %) is a good starting point for the calculation. This value was used e.g. for the calculation that produced figure 2.

Four of the DSYHS models were also tested with a different length of the transom overhang. The length of the waterline at rest is not influenced by the overhang. This offers the opportunity to test the chosen definition of the Froude number. The residuary resistance for the current definition is depicted in figure 3. The variation of the transom overhang length only plays a role at the higher speeds, when the transom is submerged for the models with the shorter overhang.

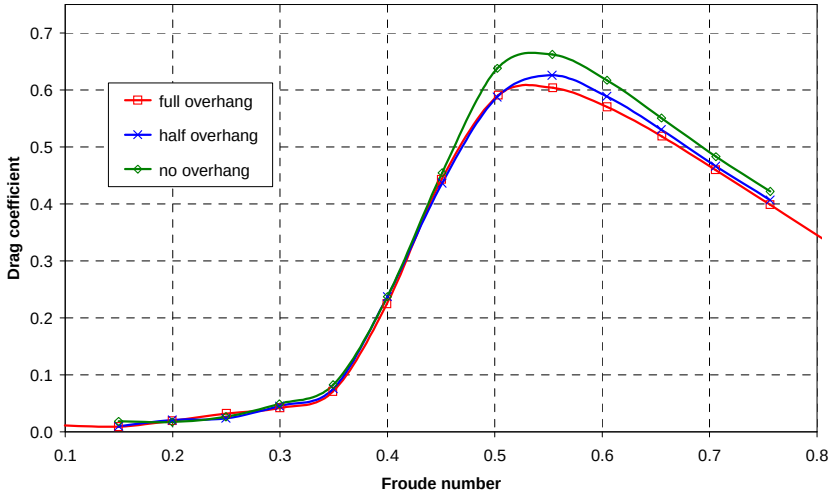

Figure 3. Residuary resistances coefficient for Sysser 27 with 3 overhang variations. Froude number based on waterline length at rest

$$C_{residu} = \frac{R_{residu}}{\frac{1}{2}\rho \cdot U_\infty{}^2 \cdot \frac{V_{CB}}{L_{WL}}}$$

Instead of using the waterline length at rest one could also base the Froude number on the actual length of the waterline, including the influence of the wave profile along the hull. The wave profiles for Sysser 26 and 27 were published in [37]. With this information the actual wetted length at speed was estimated for Sysser 27 for all three overhang variations. The length of the overhang has a significant influence on the wetted length and hence on the Froude number in the new definition. The resistance curves are shown in figure 4.

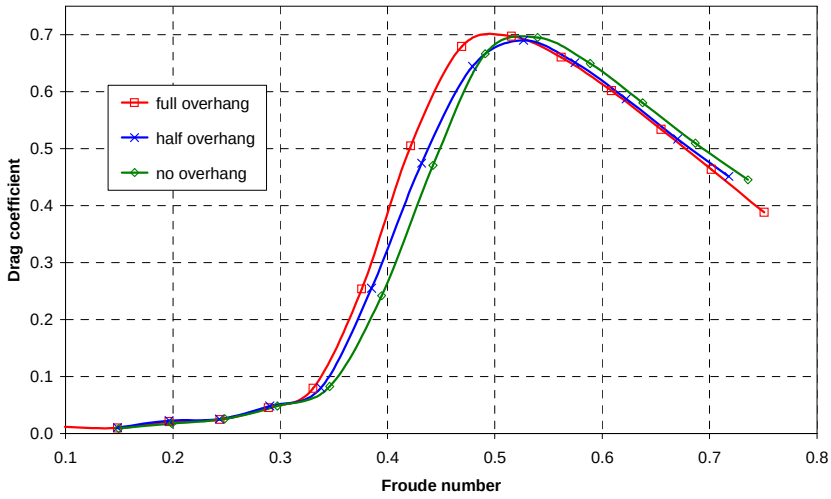

Figure 4. Residuary resistances coefficient for Sysser 27 with 3 overhang variations. Froude number based on actual wetted length at speed.

$$C_{residu} = \frac{R_{residu}}{\frac{1}{2}\rho \cdot U_\infty{}^2 \cdot \frac{V_{CB}}{L_{WLwet}}}$$

The curves are shifted and the similarity is lost. It is obvious that the wetted length at speed, including the wave profile along the hull, is not the wavemaking length that is required in the denominator of the Froude number for the sake of similarity. The initial definition of the Froude number, using the waterline at rest, was therefore considered to be a better guess for the hydrodynamic similarity and was kept throughout this report. It is also clear that the transom overhang is a relevant parameter for the residuary resistance and must be added to the list.

3. GENERATING THE DATABASE

In the next step the relevant geometric dimensions of the hull and all other properties that might influence the residuary resistance must be identified and listed. This is the most crucial part of the whole process. If an important parameter is left out in this step, the correlation can never be successful. To be on the safe side it is best to include too many parameters at this step and sort out the irrelevant ones later, based on the correlation results. To broaden the covered parameter range two additional models were added to the database. Available were tank test data of a Dehler 33 [34] and a Delft-372 [35].

3.1 Selection of the parameters

The Delft-method [1] uses the following physical quantities: R_{res}, U_∞, V_{CB}, ρ, g, L_{WL0}, B_{WL}, T_{CB}, Lcb, Lcf, A_X, A_W. A factor analysis [17] of these independent variables revealed that 75% of the variance in the variables across all models could be explained by only two factors. This indicates that the so called independent variables, that are used to predict the resistance force (the dependant variable), are statistically not independent. There is a high risk, that the chosen variables are not sufficient for the prediction of the resistance. The chosen variables have in common that they describe the hull form in a more global way, e.g. over all length of the water line and total volume of the canoe body, whereas the wave making is more linked to the detailed local properties of the hull. Manen & Oossanen [7] show four different wave systems that are generated by the local curvature of the hull at different stations. Andersson [18] and Lin et al. [19] subdivided the hull in a forward and rear part and calculated the geometric quantities separately for each part of the hull. Andersson also added several angles, measured in the plane of the cross sections and also in the water plane. Fung [20] used similar local variables although not as many and not as detailed as Andersson. The influence of an additional trimming moment was modeled at Delft [21] as a function of the height of the longitudinal metacentre. The correct quantity would have been the height of the metacentre above the centre of gravity. Unfortunately the location of the centre of gravity is not known for the models. The height above the centre of buoyancy BM is used as a substitute. Initially only the first 18 quantities in the following list were used, but for reasons that are explained in chapter 5.1 the number of parameters was not sufficient. Further parameters that might have an influence on the resistance were included. After several trial loops the list of relevant quantities is now as follows:

$$R_{res},\ U_\infty,\ V_{CB},\ \rho,\ \nu,\ g,\ L_{WL},\ B_X,\ T_X,\ Lcb,\ Lcf,\ A_X,\ A_W,\ V_1,\ L_1,\ I_E,\ E_R,\ BM,\ U_{Bow},\ \alpha,\ L_{TO},\ V_{attn}$$

The maxim values B_{WL} and T_{CB} used in [1] are replaced by the values at the maximum area section B_X and T_X to produce consistent dimensionless variables in the next step. The incidence angle at the entrance I_E is the angle between the waterline and the symmetry plane, measured in the water plane at the forward end of L_{WL}. Similarly the exit rocker angle E_R is the angle between the keel line and the water plane, measured in the symmetry plane at the aft end of the water line. α is the flare-angle at the still water level at the maximum area section. The speed U_{Bow} at the edge of the b.l. 4% behind the forward end of the L_{WL} is added because Raven [22] proposed to use the pressure coefficient of the double-body potential flow as an input to the computation of the wave resistance at low speeds. The pressure coefficient is defined as $1-(U_{Bow}/U_\infty)^2$ and the speed U_{Bow} is taken from the potential flow calculation of chapter 2 above. The variable V_{attn} is defined in the spirit of Michell's wave making theory as described e.g. in [23]. The wave resistance is a function of the hull offsets multiplied with an attenuation factor that depends on the immersion z. If A_{Wz} is the water plane area at the depth z below the still water level, then an attenuated Volume is defined by:

$$V_{attn} = \int_0^{T_{CB}} A_{Wz} \cdot e^{-z \cdot g / U_\infty^2} \, \mathrm{d}\, z \tag{8}$$

A further question that needs to be discussed is the influence of the height of the towing point above the still water level. This height equaled in the experiments a value between 9% and 20% of the L_{WL0}. The varying lever arm creates different trimming moments from model to model; the influence of this variation was not corrected in [1]. For some test runs an additional trimming moment at zero speed was applied by adjusting a movable weight on the centerline of the model. To compensate for these variations the following process is used for the new analysis: The trimming moment, which is the drag force times the lever arm of the towing point relative to the center of effort of this hydro-dynamic drag force, and any additionally applied trimming moment at rest are

added together. The sinkage and trim angle are calculated for this combined trimming moment in the hydrostatic program. All required geometric quantities are then determined for this trimmed position at zero speed. This process avoids the addition of further variables like towing height or trimming moment. The vertical position of the hydrodynamic drag force is assumed to be at the geometric center of the wetted surface. This is an acceptable approximation for the bare hull below planing speeds. On an appended hull the drag will act at a much lower position.

There are files available in [2] with the geometric offsets (surface-points) of all tested hulls. From this source it was possible after some cleaning up, to calculate the required geometric quantities for all DSYHS models.

3.2 Dimensional analysis

It is possible to reduce the 22 parameters by three through the introduction of dimensionless variables. The tool for this is the dimensional analysis. The choice of the parameter combinations is arbitrary, but it is helpful to define the variables in accordance with known theoretical results. A good dimensionless form of the dependent variable that makes use of equation 8 is:

$$Y = \frac{R_{res}}{\frac{1}{2}\rho \cdot U_\infty{}^2 \cdot \frac{V_{attn}}{L_{WL}}} \tag{9}$$

The independent variables adopt several named dimensionless parameters already in use in naval architecture. Because $L_{WL} = L_1 + L_2$, these three lengths can be interchanged in the dimensionless variables. The same applies to $V_{CB} = V_1 + V_2$. For the dimensionless volume there are three possible forms in use:

$$\frac{V_1}{L_1{}^3} \qquad \frac{V_1{}^{1/3}}{L_1} \qquad \frac{V_1}{A_X \cdot L_1}$$

The first version is the direct outcome of the dimensional analysis, the second is an often used form of lower order and the third is the prismatic coefficient. Which of these versions will give the best correlation can not be known up front. The prismatic coefficients are selected in a first try, the alternative forms will be checked in parallel. Similarly there are three forms for the dimensionless water plane area:

$$\frac{A_W}{L_{WL} \cdot B_X} \qquad \frac{V_{CB}{}^{2/3}}{A_W} \qquad \frac{V_{CB}}{A_W \cdot T_X}$$

The first version is the water plane coefficient, the second is the form used in [1] and the third version is the vertical prismatic coefficient. In addition it is possible to define split versions for the fore- and after-body separately. All in all these are 9 possible versions. In preliminary tests the quality of the correlation was checked with all 9 versions and the winner was the water plane coefficient. The complete list of the 18 independent dimensionless variables for the further analysis is:

$$C_{P1} = \frac{V_1}{A_X \cdot L_1} \quad C_{P2} = \frac{V_2}{A_X \cdot L_2} \quad C_{WP} = \frac{A_W}{L_{WL} \cdot B_X} \quad C_X = \frac{A_X}{B_X \cdot T_X} \quad LCB = \frac{Lcb}{L_{WL}} \quad LCF = \frac{Lcf}{L_{WL}} \quad \frac{B_X}{L_1} \quad \frac{T_X}{L_1}$$

$$\frac{L_2}{L_1} \quad \frac{BM}{L_{WL}} \quad Re = \frac{U_\infty \cdot L_{WL}}{\nu} \quad Fn = \frac{U_\infty}{\sqrt{g \cdot L_{WL}}} \quad I_E \quad E_R \quad C_U = \frac{U_{Bow}{}^2}{U_\infty{}^2} \quad \alpha \quad \frac{L_{TO}}{L_{WL}} \quad \frac{V_{attn}}{V_{CB}} \tag{10}$$

The three angles are already dimensionless and can be kept as they are. The pressure coefficient is designated C_U as the traditional C_P is already in use for the prismatic coefficient.

3.3 Statistical inspection of the database

Because of varying test speeds and the change of L_{WL} by the trimming moment, the resulting Froude numbers were unevenly distributed. For the statistical analysis it is required to have sample points at discrete and fixed Froude numbers. With the means of rational cubic splines a new database was interpolated for all variables (10) at a maximum of 15 fixed Froude numbers from 0.10 to 0.80. This resulted in 1250 independent datasets. For the highest Froude number there were only 15 test points available. As this number is too small for the planned regression analysis, the resistance curve of 26 models was slightly extrapolated up to Fn = 0.8. This brought the total number of datasets up to 1276.

All textbooks on statistics recommend an inspection of the database before starting the regression analysis. Special attention should be given to the problem of collinearity. This arises when some of the independent variables are almost linearly dependent. To illustrate the problem one can imagine a sample of points $(y,x_1,x_2)_i$ in space with y depending linearly (only to keep the example simple) on the variables x_1, x_2. If these points are dispersed properly, it is possible to draw a unique plane through these points. If x_1, x_2 become linearly dependent, the points are gathering along a line and the orientation of the plane in space is not robustly defined any more. Small changes in the variables (measurement errors) can even flip the plane. A measure of collinearity is the variance inflation factor VIF [24]. If it exceeds the value of 10, the regression will be affected by collinearity. Table 1 lists the results for Fn = 0.25. For other Froude numbers the results are practically identical. The first impression is depressing and many would regard the database as useless, but the case deserves a closer look.

X	Re	C_{P1}	C_{P2}	C_X	C_{WP}	LCB	LCB/LCF	B_X/L_1	T_X/L_1	T_X/L_2	I_E	E_R	α	C_U	BM/L
VIF	7.6	15.6	26.7	48.5	17.2	47.8	41.1	32.9	100.8	102.8	106.1	90.4	25.7	16.3	35.3

Table 1. Variance inflation factors for Fn = 0.25

The VIF will grow large, if the independent variable in question is collinear with just one of the other independent variables. With so many variables this risk is high. In the final regression analysis only a subset of the variables will be used and it is possible to leave redundant variables out. A source of collinearity is also the process of the dimensional analysis. A single geometric length, e.g. L_{WL} or a part of it, is used to make all variables dimensionless. As a result, all variables have a common denominator, which fosters collinearity. Another cause of collinearity is a small spread of the variables. Only half of the variations of the variables come from the large changes from one model to the other, the other half is caused by the change in the initial trim at zero speed. These trim changes produce only minor changes in the independent variables, which leads to a reduced spread.

A helpful numerical experiment was conducted by Taylor [25]. A Monte Carlo simulation produced the result that the ill-effects of collinearity become only apparent, when the coefficient of determination $\mathscr{R}^2$ between two independent variables is larger than $\mathscr{R}^2$ of the dependent variable in relation to the full set of the overall model. Table 2 was computed to test for this criterion. It displays the coefficients for the cross-correlation between 14 independent variables for Fn = 0.25. The worst case is the correlation between T_X/L_1 and T_X/L_2 with a value of r = 0.93 i.e. $\mathscr{R}^2$ = 0.865. The value of $\mathscr{R}^2$ for the complete set of the 14 variables of table 2 in relation to the dependant variable Y is 0.93 and therefore higher than the worst cross-correlation. According to Taylor this colli-

	BM/L	C_U	α	E_R	I_E	T_X/L_2	T_X/L_1	B_X/L_1	LCB/LCF	LCB	C_{WP}	C_X	C_{P2}	C_{P1}
C_{P1}	-0.16	-0.61	0.17	0.53	0.74	0.14	0.34	0.59	0.07	-0.65	0.52	-0.50	0.36	1
C_{P2}	0.04	-0.20	0.15	0.29	0.39	-0.06	-0.10	0.15	0.22	0.30	0.77	0.03	1	
C_X	0.58	0.57	0.09	-0.80	-0.75	-0.68	-0.77	-0.65	-0.71	0.55	0.12	1		
C_{WP}	-0.03	-0.33	0.07	0.23	0.41	-0.14	-0.02	0.33	-0.20	-0.13	1			
LCB	0.33	0.51	0.06	-0.42	-0.53	-0.26	-0.56	-0.64	0.11	1				
LCB/LCF	-0.47	-0.30	-0.18	0.68	0.55	0.62	0.50	0.29	1					
B_X/L_1	-0.45	-0.67	0.09	0.66	0.86	0.46	0.64	1						
T_X/L_1	-0.87	-0.80	-0.55	0.88	0.63	0.93	1							
T_X/L_2	-0.89	-0.73	-0.63	0.83	0.49	1								
I_E	-0.47	-0.73	0.03	0.84	1									
E_R	-0.77	-0.82	-0.41	1										
α	0.77	0.46	1											
C_U	0.77	1												
BM/L	1													

Table 2. Pearson's correlation coefficients at Fn = 0.25

nearity is still harmless. The situation for all other Froude numbers is similar and the collinearity seems to be tolerable.

Interesting are individual values of geometrically uncorrelated variables. It is possible to design the exit angle at the stern E_R independently from the ratio T_X/L_1. An S-form in the keel line could produce any given value. Nevertheless these variables show a high cross-correlation of r = 0.88. Obviously the designer strived for a keel

line with almost constant curvature that pleased the eye causing this correlation. It is important to know that resistance predictions for yachts outside of the DSYHS are only possible if these yachts are designed after the same rules.

Rawlings et al. [24] write that if the regression is used only for prediction it is not seriously affected by collinearity as long as the prediction is restricted to the sample *X*-space. If the observed correlation structure of the sample (here the design philosophy for the hull) is not maintained, points may be far outside the sample space, even if well within the limits of each independent variable. They also mention the drawback of collinearity: It is not possible to identify the physically "important" variables, since the values of the correlation coefficients are distorted by measurement errors. The large coefficients of opposite sign are only cancelled out in the total sum. Miller [26] believes that the problems with collinearity stem from the usage of the normal equations for the computation of the least-squares fit. This is consistent with the recommendations of Press et al. [27] who encourage the reader to use singular value decomposition instead of the normal equations to avoid unstably balanced regression parameters, which are typical for multicollinearity.

4. REGRESSION ANALYSIS

Following the statements in the last chapter we may proceed with caution and use the database for a regression analysis with the aim to develop an equation for the prediction of the residuary resistance of the hull. A first decision must be made about the inclusion of the Froude number into the regression equation. Only a few researchers [28] tried to model the humps and hollows of the speed-resistance curve, but with limited success. The majority, including the staff at Delft, developed an individual regression for each of the evenly spaced, discrete Froude numbers. This method will also be used in this report.

4.1 Theoretical background

The prediction of the dimensionless resistance Y as a function of the vector $\mathbf{X}$ of the independent variables x_j can be regarded as an approximation of the true functionality by a Taylor series expansion.

$$Y(\mathbf{X}) = Y(\mathbf{X}_P) + \sum_{j=1}^{m} \frac{\partial Y_P}{\partial x_j} \cdot \left(x_j - x_{jP}\right) + \tfrac{1}{2} \cdot \sum_{j,k=1}^{m} \frac{\partial^2 Y_P}{\partial x_j \cdot \partial x_k} \cdot \left(x_j - x_{jP}\right) \cdot \left(x_k - x_{kP}\right) + \cdots \qquad (11)$$

$\mathbf{X}_P$ is the vector of all m independent variables x_j, evaluated at a point P inside the sample space. When using this equation for prediction, the point P and in consequence also the partial derivatives at P are kept constant. Equation (11) can therefore be rewritten with new constants A_j and B_{jk}

$$Y(\mathbf{X}) = A_0 + \sum_{j=1}^{m} A_j \cdot x_j + \sum_{j,k=1}^{m} B_{jk} \cdot x_j \cdot x_k + \cdots \qquad (12)$$

In practical applications a Taylor series is often terminated after the second order terms. This works well as long as the true functionality is not $Y \sim 1/x$ with a pole close to the sample space. In this case many higher order terms are needed to model the true behavior of the function. To avoid these cumbersome higher order terms, it is better to include $1/x$ as an additional independent variable. First test indicated that list (10) can be improved by adding B_X/T_X and replacing L_2/L_1 with T_X/L_2 and C_X with $1/C_X$. It is also better to add both forms of the dimensionless Volume, the prismatic coefficients as well as the $L/V^{1/3}$. The latter can be regarded as the addition of two more reciprocal values. The number of independent variables excluding *Fn* and *Re* is now 19.

The following example serves as an explanation for the further process. Let us assume that we have only four variables $x_1 \ldots x_4$. These could be any four out of list (10). We also assume that for x_1 and x_2 the quadratic as well as the linear terms are needed for the prediction, whereas for x_3 and x_4 the linear terms are sufficient. In this case equation (12) becomes:

$$Y(\mathbf{X}) = A_0 + \sum_{j=1}^{4} A_j \cdot x_j + B_{11} \cdot x_1^{\,2} + B_{12} \cdot x_1 \cdot x_2 + B_{22} \cdot x_2^{\,2} + \varepsilon \qquad (13)$$

As this is only an approximation because we terminated the Taylor series after the first and second order term respectively, it is necessary to include an error-term ε to make the equation exact. For each fixed Froude number there are N different tank tests with either different models or different initial trims. So we have N points $(x_{i1}, x_{i2}, x_{i3}, x_{i4}, Y_i)$ in our sample. For each of these N tank tests we can apply equation (13). In the statistical nomenclature with new constants and new variables this is written as

$$Y_i(\mathbf{X}_i) = \beta_0 + \sum_{j=1}^{7} \beta_j X_{i,j} + \varepsilon_i \tag{14}$$

The new variables X_{ij} represent the linear, quadratic and mixed forms of $x_{i1}…x_{i4}$. Linear regression will determine the coefficients β_j in such a way, that the sum of squares of the error-term is minimized. The error-term is also called residual, therefore the sum is named residual sum of squares *RSS*.

$$RSS = \sum_{i=1}^{N} \varepsilon_i^{\,2} = \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{M} \beta_j X_{i,j} \right)^2 \tag{15}$$

In our example there are 8 regression coefficients β. If all the quadratic and mixed terms are included in the regression, M increases very quickly, as the number of variables x is increased. The 19 independent variables that have been defined till now would result in 210 different regression coefficients β. Since the sample size N varies between 36 and 91, only a subset of the coefficients can be used. If the number of coefficients equals the number of test points ($N = M+1$), *RSS* will be zero and the predicted Y will be identical to the measured one. This is not desirable, because the coefficients will most likely model to the greater part the measurement errors instead of the physical dependency of the resistance. The choice of M is a balance between the introduction of bias, caused by the selection process, and the "noise" in the prediction by using too many coefficients ("overfitting"). The number of coefficients is $M+1$, because the constant coefficient β_0, called the intercept, must also be determined.
Some authors recommend reducing the number of the mixed terms first, while leaving the linear and quadratic terms in the model. Fung [20] argues, that mixed ("cross-coupling") terms should only be included if "such terms can be proved to have real physical significance". How is this possible? Equation (11) is a mathematical approximation of an unknown functionality with a polynomial. The true function of Y is by all means not a polynomial and the single terms of the mathematical approximation can not necessarily be identified with hydrodynamic properties in the flow. Leaving out the mixed terms $x_i \cdot x_j$ is equivalent to the assumption that x_i and x_j are the principal directions and the curvature of Y has a maximum or minimum along these directions. This assumption is normally not justified. Huetz and Guillerm [4] use 13 parameters, only linear and quadratic but no mixed terms. A reason for this choice is not given. Sahoo et al. [29] insist that "when a regression equation has two highly correlated variables…it is wrong to include their product". This is a direct quotation from Fairlie-Clarke [30], who in turn quoted this from J.R. Scott. This quotation is correct if the two variables are fully collinear, but why then would one include a completely redundant variable into the regression at all? To the other extreme Fung [20] cites in his literature survey regression methods that used up to 53 terms, including mixed terms up to the 6th order and Lin et al. [19] use 21 linear parameters and in addition 33 higher order and mixed terms. In the following analysis it will be decided for each variable that is entered into the regression, whether its influence should be linear or quadratic. In the quadratic case the linear and the quadratic terms are entered plus all mixed terms that contain the new variable and the quadratic variables already in the regression, just as in the example equation (13).

The task is now to select a prediction model with only a few of the 210 possible coefficients. In theory there are $2^{210} = 1.6 \cdot 10^{63}$ different possible models! We therefore need a process for the selection of a good, hopefully the "best" (whatever that means) subset of variables. Variable selection in regression has received an enormous attention in the literature in the last 20 years. The state of the art in 2002 is described in [26].

4.2 Variable selection

One possibility is to select the variables manually by experience, intuition or trial and error. A first guidance for the significance of the predictor-variable is its direct correlation with the dependent variable. The correlation coefficients for some of the variables, linear as well as quadratic terms, are depicted in figure 5. A striking feature is the sign-change of the correlation coefficients between $Fn = 0.35$ and 0.45. An increase of e.g. T_X/L_2 will decrease the resistance at low Froude numbers, whereas it will increase the resistance at higher speeds. The wave system created by the hull could give an explanation for this effect. At $Fn = 0.4$ the length of the wave created by the bow equals L_{WL}, the second wave crest is therefore at the stern. At speeds below $Fn = 0.4$ the stern is supported by the second crest of the bow wave, whereas above $Fn = 0.4$ the crest is behind the ship, the stern

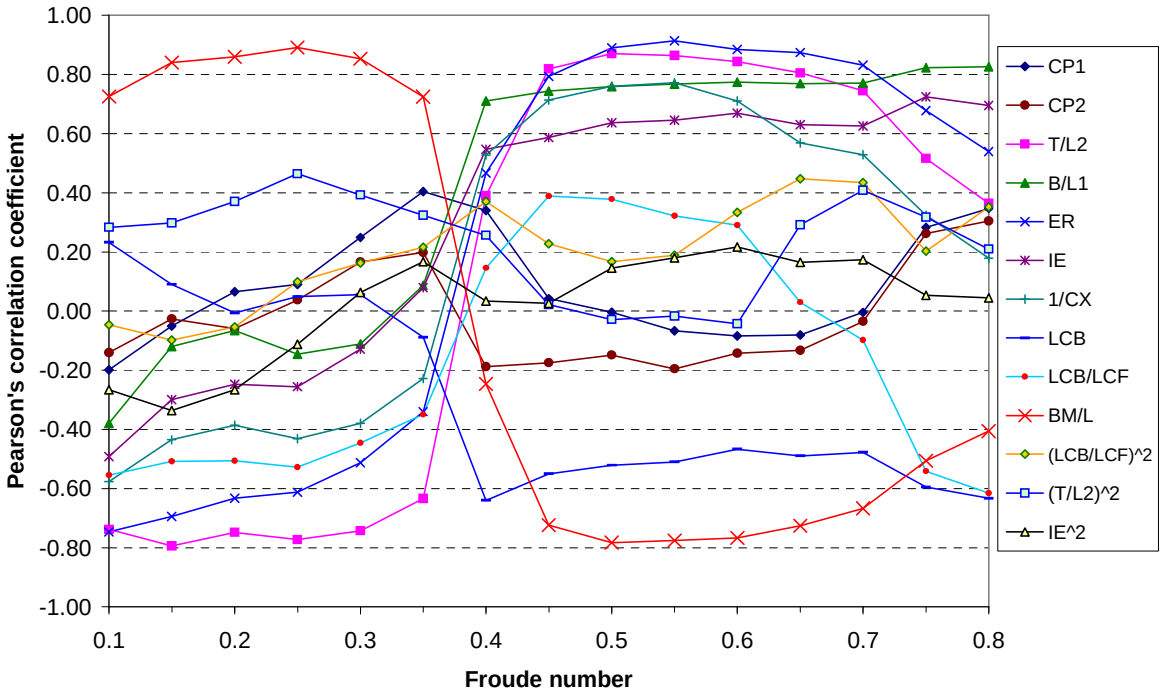

Figure 5. Correlation of the individual X_i to the Y_i

falls into the trough, the bow goes up and the ship is driven "uphill". This occurrence of speed regimes where parameter influences change might indicate the necessity to use different variables for modeling different speeds. In a first test only 17 linear variables without *Re* and B_X/T_X were included in the regression analysis. This can be regarded as the starting point of the variable selection process. The quality of fit is measured by the standard deviation of the residual sum of squares. The definition is:

$$\sigma = \sqrt{\frac{RSS}{N-M-1}} = \sqrt{\frac{1}{N-M-1}\cdot\sum_{i=1}^{N}\varepsilon_i^{\,2}} \tag{16}$$

The denominator under the root sign represents the degrees of freedom of the error vector. The *RSS* in itself is not a good measure of fit, because it decreases with each added variable till it reaches zero. The *RSS* of a subset will always be higher than the *RSS* of the complete set. The standard deviation is a better measure, because it will increase, if an additional variable only adds noise and no useful information. The target is therefore to find a subset with a low standard deviation.

The results for the regression with 17 linear variables are listed in table 3 for each Froude number. To make the values comparable, σ was standardized for each Froude number separately with the mean value of the Y_i. In a next step the quadratic and mixed terms for 4 variables were added. The selection of the quadratic variables was guided by the correlation results in figure 5. The result is listed in table 3 in the line "manual selection". It can be seen that below *Fn* = 0.25 there are too many terms, σ has increased compared to the linear case.

Searching the literature of the last decade on variable selection, there is one algorithm that stands out. Many authors recommend as best practice the use of Lasso (least absolute shrinkage and selection operator) [26]. In Lasso equation (15) is minimized for the standardized variables subject to the constraint

$$\sum_{j=1}^{M}\left|\beta_j\right| \le s \tag{17}$$

where *s* is a threshold that is gradually reduced. In each step s is adjusted in such a way, that one of the remaining constants β_j is reduced to zero and removed from the list. The results are often better than the manual selection, see table 3, but there is one big disadvantage: in a large number of cases there were mixed variables selected, without selecting the involved linear or quadratic terms which is not consistent with the requirements of equation (11).

		Fn	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80
	M+1	*N*	42	69	92	108	109	109	109	108	107	105	94	71	62	50	41
linear terms only	18	σ [%]	59.3	28.8	21.4	15.3	12.5	9.6	6.7	5.0	4.1	3.8	4.8	4.7	4.3	3.9	4.8
manual selection	28	σ [%]	62.1	31.1	22.7	15.1	10.9	8.8	5.5	4.6	4.0	3.5	4.3	4.1	2.7	3.5	4.4
Lasso	25	σ [%]	39.3	26.6	17.5	12.0	11.3	8.8	5.8	4.1	4.3	4.6	4.4	2.3	2.3	2.3	2.4
individual forward selection	21-28	σ [%]	41.4	23.2	17.3	13.0	9.5	7.6	4.2	3.8	3.5	3.4	3.9	3.6	2.8	2.4	3.7
individual full search	25	σ [%]	29.9	19.6	16.2	11.6	8.3	6.9	3.9	2.7	2.9	2.8	3.3	2.6	1.8	1.8	1.6
full search in blocks	25	σ [%]	31.3	26.7	11.7	10.2	7.8	6.6	4.5	3.0	2.8	2.6	2.9	2.5	2.5	2.0	3.4

Table 3. Standard deviation σ of the residual sum of squares in % of the mean

It would be good to have a selection process that allows at each step to control the way how the variables are included, be it linear or quadratic, including mixed terms. A method that allows this is the forward selection. This method has a bad reputation [26], which might stem from the fact that the minimum of the *RSS* is only considered for the next step. This is comparable to a chess player who makes his decisions only looking one move ahead. In reality a good player anticipates the next possible three or four moves and is guided in his decision by the analysis of all possibilities. A method was therefore programmed, that extends forward selection to the analysis of all the next possible four selections. For the initial set of 12 variables one has the choice at each step between 12 linear variables and 12 sets of quadratic and mixed terms. For each of these 24 choices there are again 24 possible choices in the next step and so on, all together $3.3 \cdot 10^5$ possibilities. Out of all these possibilities that path is identified that gives in the end the smallest standard deviation. This path consists of four variables, either in the linear or quadratic form. Out of these four variables the one is selected that allows the largest reduction of σ in the first step. After this selection, the whole process is repeated, always looking four steps ahead. The results of this stepwise regression are depicted in figure 6. A different behavior is again visible between the speeds below $Fn = 0.4$ and above. The first coefficient on the abscissa represents β_0, which is just the mean of all experimental values Y_i at that specific Froude number. If we look at $Fn = 0.15$, the standard deviation of the distribution of the Y_i around this mean is 79%. By introducing the dependence on the hull parameters it is possible to reduce σ down to 23%. The optimum (= minimum) is reached at $M+1 = 23$. If we look at $Fn = 0.4$, the picture is quite different. The standard deviation of the mean is only 14% and can be reduced by the inclusion of the hull parameters down to 4.8% for $M = 20$. So one can notice that at $Fn = 0.4$ the resistance is almost fixed and nearly independent of the hull parameters. At higher speeds the picture is again similar to the low ones, e.g. at $Fn = 0.5$ the mean is 29% and it can be reduced to 3.5% at $M = 21$. The results of the stepwise forward regression are also included for comparison in table 3. This method yields slightly better results than the Lasso but, more important, avoids the drawback of the inclusion of mixed terms without the corresponding linear and quadratic terms.

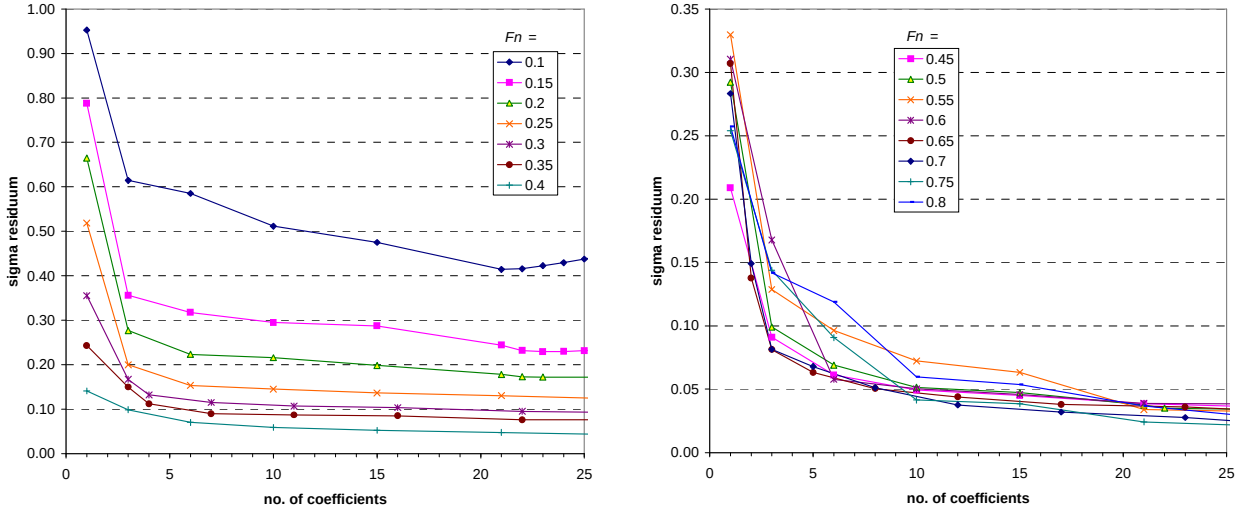

Figure 6. Standard deviation σ as a function of the number of coefficients (= *M*+1)

A problem that has not yet been discussed is the stopping criterion. In a search of the naval literature on regression analysis it was found that authors e.g. [20], [31] and [32] had only intuitive and arbitrary criteria. The decision on how many variables to include and when to stop is the balance between under- and overfitting. Burnham and Anderson [33] explain that an underfitted model is too simple, has large bias and fails to identify effects that are supported by the data. An overfitted model includes spurious variables and errors due to random fluctuations (measurement errors). According to them, underfitted models present a more serious issue in variable selection than overfitted models. In the previous paragraphs the change in σ was used as a stopping criterion, but it gives no sharp cut-off and the minimum occurs sometimes at astonishing high values of *M*. A stopping criterion that is especially suited to cases where *M* is not small compared to *N* is the second order variant of Akaike's information criterion [33]:

$$AIC_C = N \cdot \ln\left(\frac{1}{N} \cdot \sum_{i=1}^{N} \varepsilon_i^{\,2}\right) + 2K + \frac{2K \cdot (K+1)}{N-K-1} \qquad \text{with} \qquad K = M+2 \tag{18}$$

Miller [26] reports a form where $K = M+1$, but the differences are small. The relative AIC_C values for the points of figure 6 are shown in figure 7. The optimal *M* is the minimum of AIC_C. With the exception of the special case $Fn = 0.4$, the optimal number of coefficients is always less or equal 25. A closer look at the selected variables shows that the optimum contains always 5 quadratic forms including the 5 linear variables and additionally one to four other linear variables.

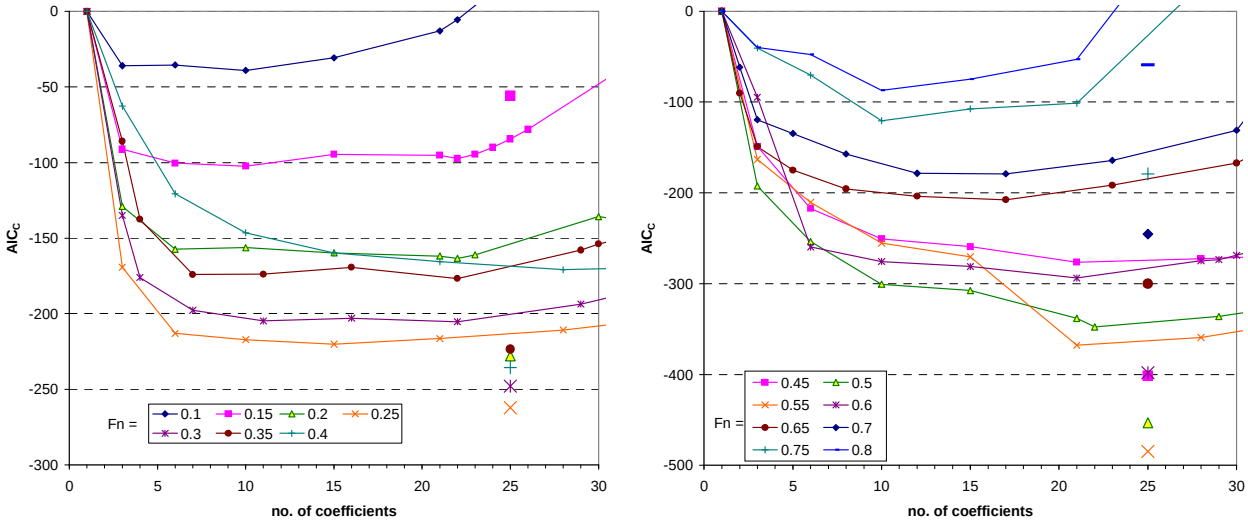

Figure 7. AIC_C as a function of the number of coefficients (= *M*+1)

All books on statistics agree that the only sure way of finding the best fitting subset is the exhaustive search. With the knowledge from the forward selection process as described in the last paragraph it is now possible to limit the complete search to a manageable size. If we chose out of 19 variables 5 variables with quadratic and mixed terms and 4 variables with linear terms only, the number of coefficients is limited to 25. In this case there are $11.6 \cdot 10^6$ models per Froude number. A complete inspection of all these models for all Froude numbers can be performed within a week on a modern PC. During this analysis some peculiarities could be observed.

- The choice of additional variables depends on the number of coefficients. Let us assume that besides X_{i1} to X_{i4}, X_{i5} is a variable with a good correlation, but the combination of X_{i6} and X_{i7} models the influence even better, then X_{i1} to X_{i5} will be chosen if the number of coefficients is restricted to 5. If 6 coefficients are allowed X_{i5} will be dropped and X_{i6} and X_{i7} will be chosen instead. It is therefore almost impossible to recognize a pattern or a ranking of the variables in the choosing process.
- Sometimes it was possible to get the same standard deviation σ with different sets of variables. This is a sign that there are redundant variables in the list. The full search with a tightly limited number of coefficients avoids the inclusion of redundant variables.
- Often a good correlation could only be achieved for part of the experiments. One set of variables gave a good correlation for the old models of series 1, but did not predict well the newer models and a different set of variables predicted the new models well, but not the old ones. This was taken as an indication that more than 25 coefficients are needed in a set.

- When more than 25 coefficients were used, the prediction was not as robust as needed. This became visible when the height of the towing point was varied. As soon as the limits of the DSYHS sample space were surpassed the predicted curve of resistance vs. Froude number exhibited strong oscillations.
- The AIC_C values were a good guidance to distinguish between competing models with a different number of coefficients when the difference in the standard deviation σ was small.

These problems with conflicting demands are not solvable. The compromise that was chosen first, allowed different variable selections for different Froude numbers. The results of the full search can be found in table 3 in the line "individual full search". They were the best for the initial set of variables. This selection together with Lasso and the individual forward selection have a big disadvantage in common; they use different sets of variables for each Froude number. From a hydrodynamic point of view this is questionable. The selection will most likely be driven by measurement errors. The resulting correlation equation predicts the measured values very well, but it fails, when the height of the towing point is varied or when it is applied to new models, not in the DSYHS. In such cases the predicted curve of resistance vs. Froude number is wavy and not smooth.

In the next step the full search was therefore programmed to find the optimum set for a range of consecutive Froude numbers. The full range from 0.1 to 0.8 was divided into 3 blocks, guided by figure 5. For each block a separate set of variables was determined. The program searched in each block for the parameter set that gave the smallest sum of σ^2. The resulting standard deviations are listed in the last row of table 3. This time the new database with 19 parameters was used. The AIC_C values of the selected regressions are entered as single large symbols at no. 25 into figure 7 for comparison. The results with the new parameters are as good as or even better than in the case of the individual search. The subsets that were finally chosen are com-piled in the upper half of table 4.

	fore body						*after body*					*entire hull*					*max. area section*		
Fn	C_{P1}	$L_1/V_1^{1/3}$	T_X/L_1	B_X/L_1	i_E	C_U	C_{P2}	$L_2/V_2^{1/3}$	T_X/L_2	E_R	L_{TO}/L_{WL}	V_{attn}/V	*LCB*	*LCF*	*BM/L*	C_{WP}	$1/C_X$	B_X/T_X	α
0.1 – 0.35		X		X	I	I	I						X			X	X		I
0.4 – 0.5				I	I	I		X	X	X	X	X	I						
0.55 – 0.8				I					X	X	X	X	I	I	X			I	
outside of the DSYHS																			
0.1 – 0.3	I	I	I	I		I				I	I	I	I	I		I	I	I	
0.35 – 0.45	I		I				I	I	I	I	I	I	I	I		I	I	I	
0.5 – 0.8	I			I			I		I	I	I	I	I	I	I	I	I	I	

I = linear term only X = linear, quadratic and mixed terms

Table 4. Finally selected variables

The choice of the Froude number ranges is arbitrary and influences the results of the full search. As already indicated, the collinearity of the variables does not allow the identification of the physically "right" parameters, they are disguised in the "noise" created by the measurement errors. Many more models, a much larger sample size and most likely smaller measurement errors would be needed to identify the variables that really determine the wave resistance. The current selection does not render the prediction invalid, one just has to accept, that the uncertainty of the measurements is not averaged out, but is contained in the predicted resistance.

4.3 False friends

At first sight the Reynolds number is closely correlated with the resistance. Computing the probability of the null hypothesis by a comparison with Student's *t*-distribution gives a zero probability for the case that no correlation exists between *Re* and *Y*. Numerical values are listed in table 5. In addition table 1 shows that *Re* is the only variable that is not strongly correlated with other variables. So *Re* seems to be an ideal predictor variable. On top it is not difficult to find physical arguments that viscosity should have some influence on the residual resistance. For all these good reasons *Re* was included in the regression analysis and the linear term was kept in the final set of predictor variables. When comparing the predicted resistance with the measured values of the DSYHS, the

agreement was very good. The surprise came, when a comparison was made with tank test data for a model of 5 meter length [34]. The maximum length of the DSYHS models is 2 meter, so there is a substantial extrapolation necessary for this prediction. The results were hilarious, extrapolation was not possible. An inspection of the raw data was the only way how to find the root cause of this false regression model. Figure 8 shows experimental data for $Fn = 0.25$.

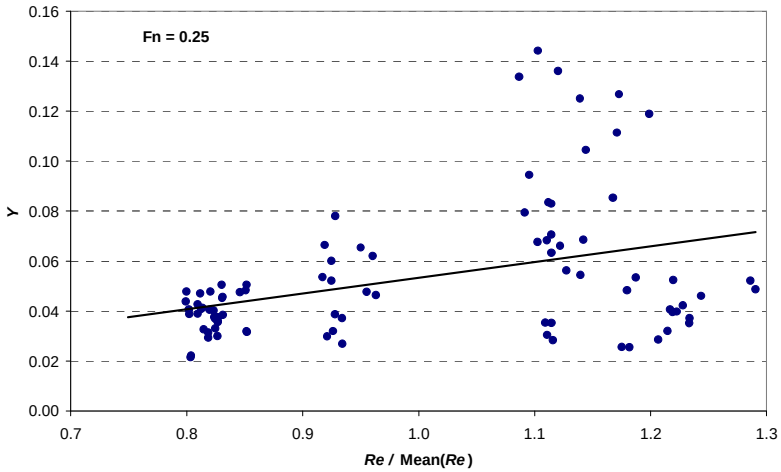

Figure 8. Correlation between dependent variable *Y* and standardized Reynolds number at Fn = 0.25

This Froude number is picked, because the results are typical. Several different model families were used in the DSYHS with three different model lengths. The length was kept constant within the family; therefore the points are clustered around 3 Reynolds numbers. The family with the short model was the old design with a V-shaped forebody, whereas the longer models had a modern dinghy-like hull. So it is clear that the model length was not chosen randomly. The variation of the Reynolds number was not extended across all the different hull forms and can therefore not be used for the regression. As a consequence the whole analysis as described in the previous chapters had to be repeated, this time without *Re* in the subsets. The standard deviations σ were equally as good as before. To make sure that the drop of *Re* from the variable list was right, a regression with the final sets of table 4 was performed. The predictions were calculated and then the residuals ε_i, as defined in equation (14), were correlated against *Re*. Results are in table 5.

	Fn	0.10	0.15	0.20	0.25	0.3	0.35	0.40	0.45	0.50	0.55	0.6	0.65	0.70	0.75	0.80
Correlation with *Y*	$\mathscr{R}^2$	0.311	0.149	0.076	0.125	0.097	0.028	0.358	0.545	0.614	0.638	0.617	0.587	0.586	0.604	0.618
	Probabil.	0	0	0.02	0	0	0.12	0	0	0	0	0	0	0	0	0
Correlation with ε	$\mathscr{R}^2$	0	0.003	0.001	0	0.003	0.006	0.002	0.004	0.001	0.002	0.003	0.001	0	0	0
	Probabil.	0.97	0.70	0.85	0.99	0.59	0.45	0.70	0.56	0.78	0.68	0.62	0.80	0.90	0.96	0.98

Table 5. Correlation of *Re* against the measurements and against the residuals

The coefficients of determination are all close to zero. The values of the probability are again for the hypothesis that there is no correlation. It shows that with a probability between 45% and 98% a random selection would yield the same correlation. So in this case it is proven, that the residual errors do not correlate with *Re* and it is absolutely right to leave *Re* out. This is a classical example of selection bias during the design of the experiments.

5. PREDICTED VS. EXPERIMENTAL RESISTANCE

5.1 At model size

At the end of his publications Prof. Keuning usually compares measured and predicted resistance curves to give an impression of the quality-of-fit. The chosen examples are always the parent models of the DSYHS without additional trimming moment [1]. If the same comparison is made for the new regression with the variables of table 4, the difference between Keuning's prediction results und the new formula is only small and the additional effort of the new method would not be justified. The picture changes, when the complete database of all 1365 test points is inspected. Figure 9 compares the predicted and experimental values of the total resistance for the new regression and for the Delft-method with the addition of the influence of the trimming moment as described in [1] and [21]. The Delft-method uses all in all 14 regression coefficients. Y_{tot} in the diagrams is defined in equation 19. R_{tot} is either the measured total resistance or the sum of viscous and residual resistance.

$$Y_{tot} = \frac{R_{tot}}{\frac{1}{2}\rho \cdot U_{\infty}{}^{2} \cdot \frac{V_{CB}}{L_{WL0}}} \tag{19}$$

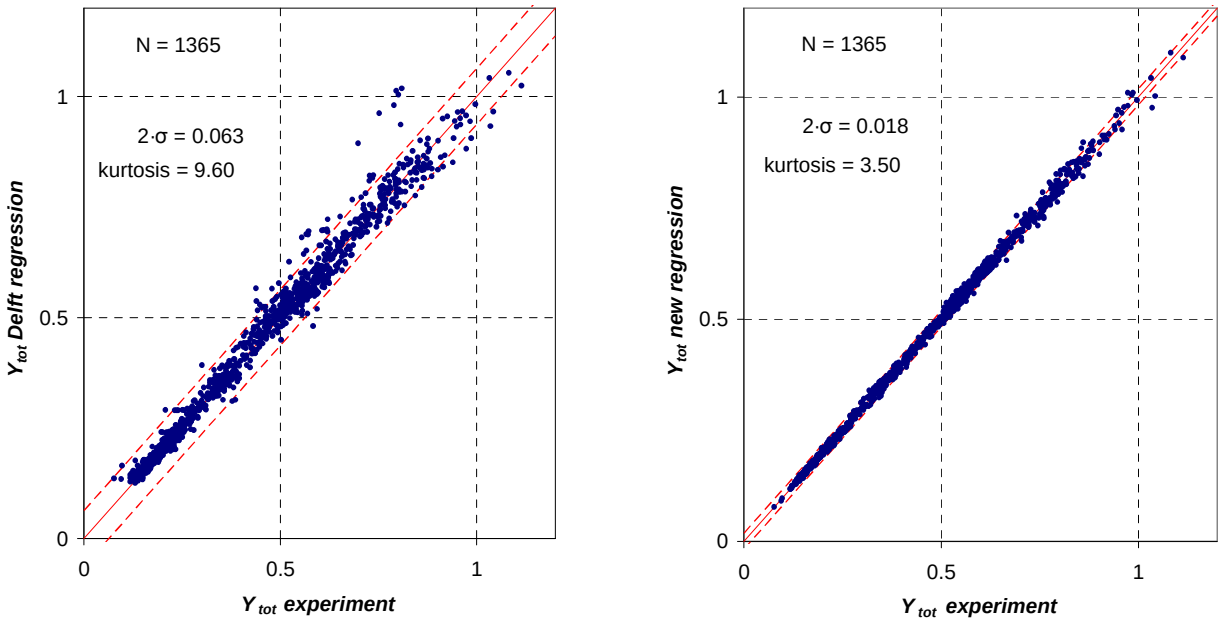

Figure 9. Total resistance coefficients for all experiments 0.1 < *Fn* < 0.77

In the ideal case all points would be lying on the diagonal line. A high value of the kurtosis is a clear sign that the distribution is peaked and not normal (Gaussian). Initially the kurtosis of the new regression was almost as high as for the Delft-method. In such a case there are many predictions very close to the experimental value but also many points in the wide tails of the distribution where the prediction is far off. This led to the conclusion that there must be additional effects that are not modeled by the initial set of variables. The process was therefore restarted in chapter 3.1 with the definition of new additional variables. The right diagram in figure 9 is the result for the final variable selection of table 4 with significantly reduced standard error and also kurtosis. The red dashed lines in figure 9 indicate the 2·σ bands. In case of a normal distribution of the errors, the ± 2·σ band would contain 95% of all test points. Because of the high kurtosis the distribution is not normal. With a sample size of 1365 it is possible to determine the quantiles empirically by counting. Astonishingly the results (95%) are in both cases identical to the values of the normal distribution. The statistic evaluation shows that in the new formula the standard deviation and also the error band are reduced by a factor of 3.5 compared to the Delft-method. Helpful is also a look at the relative error of the predicted total resistance in figure 10.

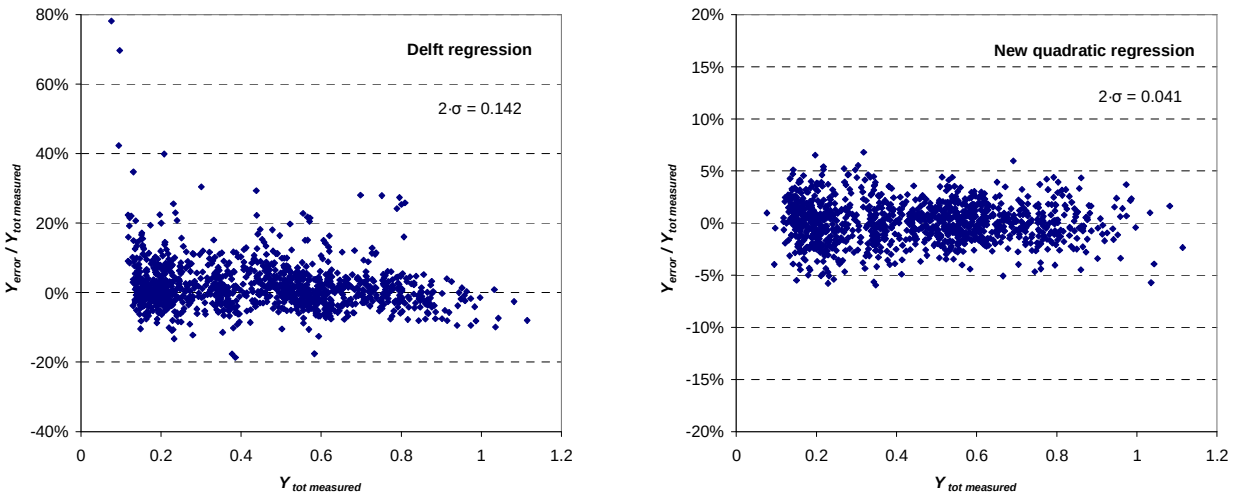

Figure 10. Total relative resistance error for all experiments 0.1 < *Fn* < 0.77

The prediction error is calculated from

$$Y_{error} = \quad Y_{tot}\,predicted - Y_{tot}\,measured.$$

It is reduced with the new method from 14.2% to 4.1%. The relative error is obviously larger at the low end for small *Fn*. This trend is more pronounced in the Delft regression than in the new one. The reduction of the errors is significant but it must be pointed out, that these errors are only valid for models of the DSYHS. For models outside of this database the error will be larger because of the selection bias in the regression analysis.

The kurtosis of the final selection is still high and the distribution is not normal. It is therefore interesting to have a look at the statistical distribution of the prediction errors for the new regression. The deviation from the normal distribution with a peak at the mean and missing values at medium distance to the mean is clearly visible in figure 11. Another way to look at this picture is the assumption of a normal distribution caused by the measurement errors and on top a superposition of a random error caused by an unknown parameter. Figure 12 depicts this hypothesis by assuming a smaller standard deviation for the normal distribution.

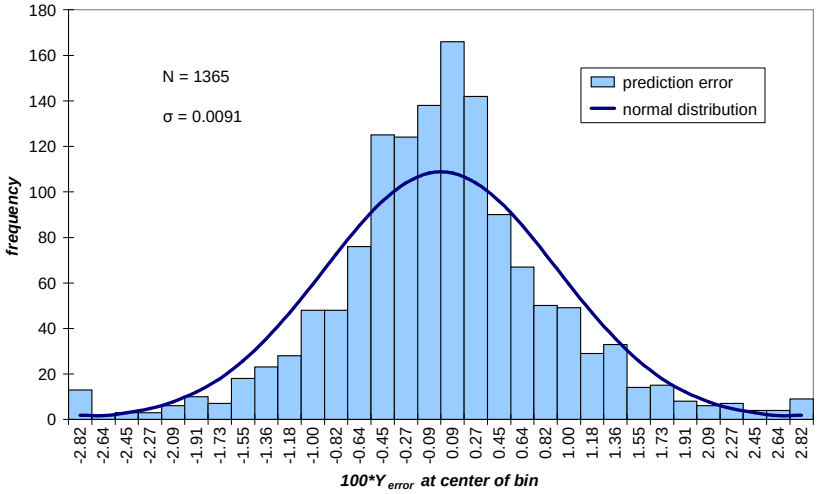

Figure 11. Distribution of the prediction error and comparison with normal distribution of equal standard deviation

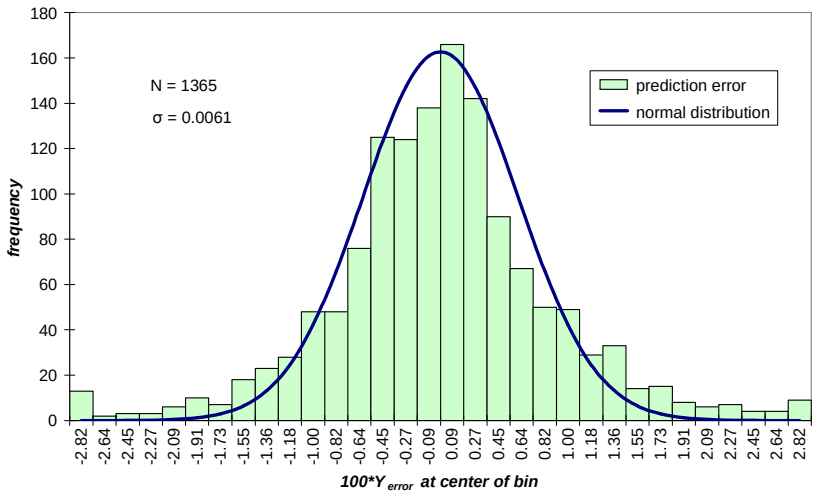

Figure 12. Distribution of the prediction error and comparison with a normal distribution of reduced standard deviation

The unknown parameter that causes additional randomly distributed small and large errors is not necessarily a hull parameter. It could also be the influence of the towing tank set up. The roughness strips are changed between the test runs, this could have an influence. Normally the measured resistance is time dependent and an averaging process is employed, which is unknown. Without a detailed knowledge of the experimental process and ideally the inspection of the raw data, a further analysis is not possible and the current error must be accepted.

5.2 Extrapolation to full size

It was already mentioned that the viscous resistance and the wave resistance are scaled differently when extrapolating to full size. A sensitive example is the model Sysser 26, because the viscous resistance according to the Delft-method differs quite a bit from the result of the boundary layer calculation. The comparison at model

size is depicted in figure 12 on the left side. The drag coefficient has the definition of equation 6. There is an agreement between predicted and measured resistance. At full size the total drag calculated with the Delft-method is about 20% lower than the new prediction, based on the b.l.- calculation, depicted on the right side of figure 12. The different total drag coefficient at full size leads with a scaled towing height also to a different trimming moment which in turn results in a new residuary drag coefficient. This change in trim when extrapolating to full size is not considered in the Delft-method. If the new prediction is valid is an open question, as long as there are no towing results at full size available. Right now it is a better founded guess than the Delft-prediction.

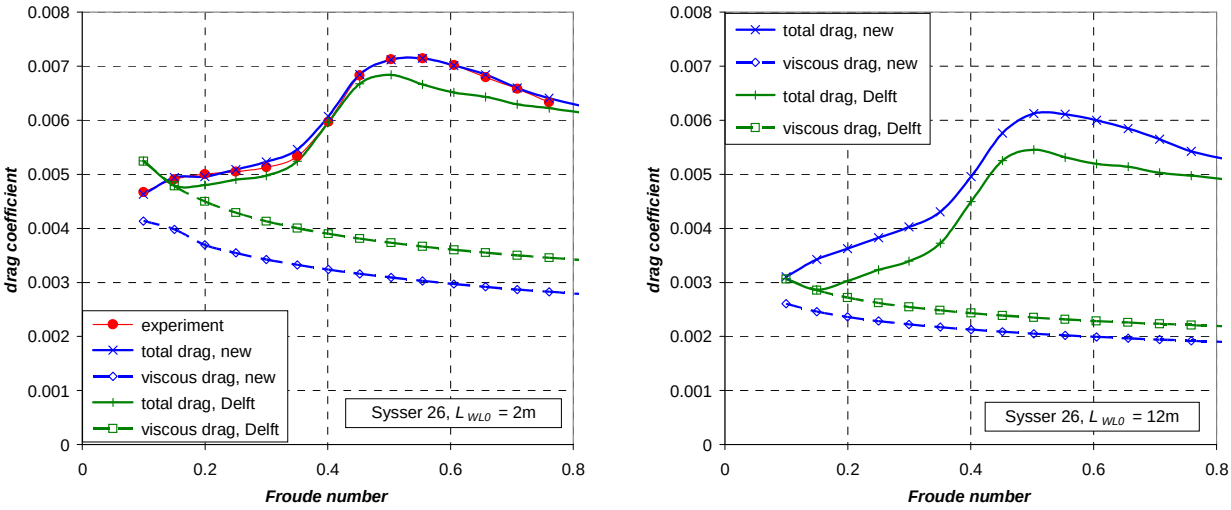

Figure 12. Drag coefficients for model and full size

6. HULLFORMS OUTSIDE OF THE DSYHS SAMPLE SPACE

At the end of chapter 3.3 it was explained that the predicted resistance can be far off, if the hull parameters and the design philosophy are not covered by the range of the tested models in the DSYHS. In such a case an extrapolation of the regression is required. Keuning and Katgert [1] deleted all higher order terms in their regression because, as they stated, these terms weakened the robustness and stability if the prediction has to be made outside of the parameter range. Following this advice an additional regression analysis was performed, this time limiting the parameters to the linear terms only. The number of variables was chosen with the help of the AIC_C. The selected variables are indicated in the lower half of table 4. The resulting standard deviations are listed in table 6.

		Fn	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80
	M+1	*N*	42	69	92	108	109	109	109	108	107	105	94	71	62	50	41
linear full search, blocks	14	σ [%]	49.3	34.3	21.1	12.3	10.1	8.2	6.6	5.1	4.3	4.0	4.6	4.4	4.6	3.6	4.4

Table 6. Standard deviation σ of the residual sum of squares in % of the mean

These deviations are higher by an average factor of 1.5 than the values for the regression that includes the quadratic and mixed terms. It is advisable to use both predictions in parallel and check them for plausibility if a hull form does not fall clearly into the range of the DSYHS models. For the models of the DSYHS the new linear regression is still better by a factor of 2 than the initial Delft–method, as can be seen when comparing figure 13 to figure 10.

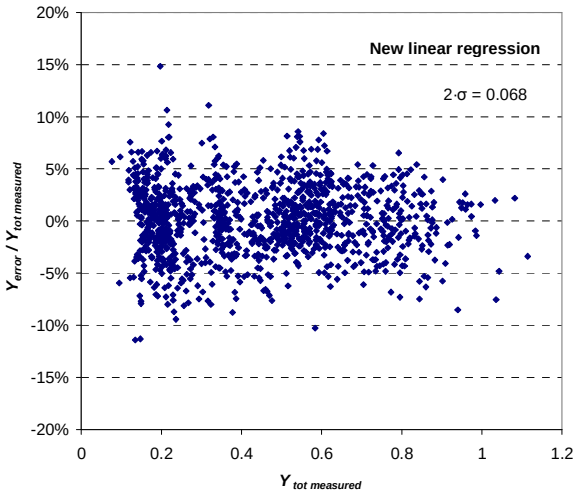

Figure 13. Total relative resistance error for all DSYHS experiments 0.1 < *Fn* < 0.77 new linear regression

7. CONCLUSION

The new regression model improves the prediction of the bare hull upright resistance compared to the Delft-method for models within the DSYHS. Future comparisons will tell, if the improvements will consistently appear also with different and new designs. To enable this necessary feed-back, the new prediction-software "UliTank" was developed and is available online.

8. REFERENCES

1. Keuning, J. A. & Katgert, M., "A Bare Hull Resistance Prediction Method Derived from the Results of the Delft Systematic Yacht Hull Series Extended to Higher Speeds", *Innovation in High Performance Sailing Yachts*, Lorient, France, 2008, RINA, pp 13-21.
2. http://dsyhs.tudelft.nl
3. Battistin, D., Peri, D., Campana, E.F., "Geometry and Resistance of the IACC Systematic Series 'Il Moro di Venezia' ", *The 17th Chesapeake Sayling Yacht Symposium*, Annapolis, Md., 2005, pp 33-51
4. Huetz, L., Guillerm, P.E., "Database Building and Statistical Methods to Predict Sailing Yachts Hydrodynamics", in *Proc. 3rd Int. Conf. Innovation in High Performance Sailing Yachts*, Lorient, France, 2013, pp. 23-32
5. van Mierlo, K. "Trend validation of SHIPFLOW based on the bare hull upright resistance of the Delft Series", M.S. thesis, Dept. Aerodynamics, Delft Univ. of Technology, Delft, NL, 2006
6. Wassink, J., "Bible for Boats", *Delft outlook, Magazine of Delft University of Technology*, No. 1, 2013, pp 6-9
7. van Manen, J.D., van Oossanen, P., "Resistance" in *Principles of Naval Architecture*, 2nd ed., Lewis, E.D., Ed., Jersey City, NJ: SNAME, 1988
8. Cebeci, T., Cousteix, J., *Modelling and Computation of Boundary-Layer Flows*, Heidelberg, Germany: Springer, 2005
9. Landweber, L., "The axially symmetric potential flow about elongated bodies of revolution", The David W. Taylor Model Basin, Washington, D.C., 1951
10. Granville, P.S., "The Calculation of the Viscous Drag of Bodies of Revolution", The David W. Taylor Model Basin, Washington, D.C., 1953
11. Patel, V.C., Guven, O., "Importance of the Near Wake in Drag Prediction of Bodies of Revolution", *AIAA Journal*, Vol. 14, No. 8, 1976, pp 1132-1133
12. Hoerner, S.F., *Fluid-Dynamic Drag*, Midland Park, NJ: published by the author, 1965
13. Jasper den Ouden, personal e-mail, 2013
14. Remmlinger, U., "The Influence of Sand Grain Strips on Boundary Layer Transition in the Towing Tank", [Online]. Available: www.remmlinger.com/B-L-trip.pdf

15. Remmlinger, U., "The Factors that Trigger the Transition to Turbulent Flow on the Keel of a Sailing Yacht", [Online]. Available: www.remmlinger.com/TurbLevel.pdf
16. Ferguson, J.M., Comment in "Subject 4–Turbulence Stimulation in Model Test", in *Proc. 6th Int. Conf. Ship Tank Superintendents*, Washington, D.C., 1951, p. 134
17. Costello, A.B., Osborne, J.W., "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis", *Practical Assessment, Research & Evaluation*, Vol. 10, No. 7, 2005
18. Andersson, G.O., "Statistische Analyse des Widerstandes von Schiffsmodellen", *Schriftenreihe Schiffbau*, No. 365, Technische Universität Hamburg-Harburg, Hamburg, Germany, 1976
19. Lin, C-W., Day, W.G., Lin, W-C., "Statistical Prediction of Ship's Effective Power Using Theoretical Formulation and Historic Data", *Marine Technology*, Vol. 24, No. 3, 1987, pp. 237-245
20. Fung, S.C., "Resistance and Powering Prediction for Transom Stern Hull Forms During Early Stage Ship Design", *SNAME Transactions*, Vol. 99, 1991, pp. 29-84
21. Keuning, J.A., Sonnenberg, U.B., "Approximation of the Hydrodynamic Forces on a Sailing Yacht based on the DSYHS", *HISWA Symposium*, Amsterdam, NL, 1998
22. Raven, H.C., "Validation of an approach to analyse and understand ship wave making", *Journal of Marine Science and Technology*, Vol. 15, No. 4, 2010, pp. 331-344
23. Tuck, E.O., "The wave resistance formula of J.H. Michell (1898) and its significance to recent research in ship hydrodynamics", *J. Austral. Math. Soc. Ser. B(30)*, 1989, pp. 365-377
24. Rawlings, J.O., Pantula, S.G., Dickey, D.A., *Applied Regression Analysis*, New York, NJ: Springer, 1998
25. Taylor, L.D., "A Note on the Harmful Effects of Multicollinearity", Univ. of Arizona, [Online]. Available: https://econ.arizona.edu/docs/Working_Papers/2012/Econ-WP-12-04.pdf
26. Miller, A., *Subset Selection in Regression*, 2nd ed., Boca Raton, FL: Chapman & Hall/CRC, 2002
27. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., *Numerical Recipes*, Cambridge University Press, 1994
28. Holtrop, J., "A statistical analysis of performance test results", *International Shipbuilding Progress*, Vol. 24, No. 270, 1977
29. Sahoo, P.K., Peng, H., Won, J., Sangarasigamany, D., "Re-evaluation of Resistance Prediction for High-Speed Round Bilge Hull Forms", in *Proc. 11th Int. Conf. Fast Sea Transportation*, Honolulu, HI, Amer. Soc. Naval Engineers, 2011, pp. 311-319
30. Fairlie-Clarke, A.C., "Regression analysis of ship data", *Int. Shipbuilding Progress*, Vol. 22, 1975, pp. 227-250
31. Mercier, J.A., Savitsky, D., "Resistance of transom-stern craft in the pre-planing regime", Stevens Inst. of Technology, Hoboken, NJ., 1973
32. Helmore, P.J., Swain, P.M., "Regression Analysis of the Ridgeley-Nevitt Trawler Series Resistance", *Proc. Pacific Int. Maritime Conference*, Sydney, Australia, 2006
33. Burnham, K.P., Anderson, D.R., *Model Selection and Multimodel Inference*, rev. ed., New York, NY: Springer, 2002
34. Fröhlich, M., "Optimierung von Kielen einschließlich Rumpf für Segelyachten auf der Basis eines numerischen Rechenverfahrens für viskose und instationäre Strömung", Schiffbau-Versuchsanstalt Potsdam, Germany, 1997
35. Remmlinger, U., " The Resistance of the Delft 372 Hull", [Online]. Available: www.remmlinger.com/Delft372.pdf
36. Allroth, J., Wu, T.-H., "A CFD Investigation of Sailing Yacht Transom Sterns", M.S. thesis, Dept. Shipping and Marine technology, Chalmers Univ. Technology, Göteborg, Sweden, 2013, [Online]. Available: http://publications.lib.chalmers.se/records/fulltext/203795/203795.pdf

37. de Ridder, E.J., Gaillarde, G., van Walree, F., "Advanced and Future Hydrodynamic Optimisation Tools in Sail Yacht Design", *2 Simposio International de diseno y produccion de yates de motor y vela*, Madrid, Spain, 2006